AECC

AUTOMOTIVE EDGE
COMPUTING CONSORTIUM

AECC Technical Report v3.0
Automotive Edge Computing Consortium (AECC)
Technical Solution Working Group (WG2)

Driving Data to the Edge: The Challenge of Data Traffic Distribution

Contents

Executive Summary

Connected vehicles are swiftly transforming the automotive industry, with emerging services driving the requirement for high volumes of data communication. New connected vehicle services are expected to give the automotive industry the fastest-growing demand for mobile connectivity. In principle, every new vehicle that is manufactured will be continuously connected; it will also generate massive volumes of data that need to be transmitted to the cloud.

Considering the global distribution of vehicle fleets and therefore the global nature of this challenge, stakeholders, including vehicle manufacturers, technology solution vendors, network operators and cloud infrastructure and service providers, should consider how communication networks can be optimized securely and cost-effectively to enable these new mobility services.

The members of the Automotive Edge Computing Consortium (AECC) are working to articulate an architecture capable of addressing the critical industry needs. To address this challenge, the AECC has proposed a “Distributed Computing on Localized Networks” solution concept and architecture that will offer the service flexibility and efficiency required to support the evolution of the automotive industry into and beyond the connected vehicle era. The solution consists of three main aspects: localized networks, distributed computing and local data integration. This technical report focuses on eight of the key issues and corresponding solutions for the localized networking aspect. Specifically, these key issues are:

Edge Data Offloading. How cellular networks can support the offloading of data from the cellular network to appropriate localized distributed computing infrastructure in an efficient and flexible manner, considering the mobility of vehicles and service requirements.

Mobility Service Instance Selection. How connected vehicles can select mobility service instances in the distributed computing infrastructure.

Vehicle System Reachability. How connected vehicles can be awakened and contacted, despite vehicle mobility and network topology changes.

Access Network Selection. How connected vehicles can select appropriate access networks according to the service requirements and capabilities of the access networks.

Provisioning and Configuration Update. How configuration parameters and policies can be provisioned to connected vehicles in a dynamic and distributed environment.

Opportunistic Data Transfer. How cellular network resources can be used to offer latency-tolerant data transfer with minimal interference to existing services.

Service Continuity. How mobility services are properly handled according to the diverse requirements on continuity when both a mobility services instance switch and an IP anchor handover happen.

Geolocation Services. How to efficiently solve the geolocation-associated mobility services due to the dynamics in which vehicles move across an edge in different geolocations.

To help stakeholders consider how to address these issues, the AECC has identified and evaluated a range of potential solutions, with this technical report containing the resulting recommendations.

The AECC continues to work on additional key issues and solutions relating to data transfer between vehicle and cloud, with further studies to be released in the future. The AECC welcomes feedback on the findings and recommendations contained within this report and will take that feedback into consideration in its ongoing work.

Terms and Abbreviations

Terms:

Access Network
The network used to connect to the immediate service provider. It may contain a WLAN, cellular network and/or wired network.

Cellular Network
A network that is typically defined by 3GPP and GSMA and operated by an MNO.

Center Server
A hardware and/or software platform to host center mobility services.

Cloud
A logical, location-independent computing platform that hosts services to store, manage and process data and which is implemented on a set of remote servers.

Computing Infrastructure
The resources and services on which other systems and services are built.

Computing Facility
A physical facility that hosts computing infrastructure and related resources (including computation, network and storage, power, cooling, etc.).

Connected Vehicle
A network-attached vehicle that shares data with other network-attached devices and servers.

Distributed Computing
A computing paradigm that divides a problem into many tasks that can be addressed by many computers.

Edge Computing
A type of distributed computing system where applications, memory and processing power are allocated to other computers in order to provide desired service levels.

Edge Server
A localized hardware or software platform that hosts edge mobility services.

Enterprise Network
A network connecting center and edge servers for a specific enterprise.

Intelligent Driving
A service that augments an Advanced Driver Assistance System (ADAS) or an automated driving system with strategic decisions based on predictions of conditions along route alternatives that are gathered using connectivity to external sources.

Local Data Integration
A platform that integrates data from localized networks and the distributed computing system.

Localized Network
A local network that covers a limited domain such as a defined geographical area.

Mobility as a Service
Integration of various forms of transport services into a single mobility service accessible on demand.

Mobility Service
A service provided to the passengers, the drivers or the vehicle manufacturers (e.g., telematics, traffic, map, car/ride sharing, insurance, etc.).

Mobility Service Instance
An application instance that implements some (or all) of a mobility service.

Mobility Service Provider
A platform-independent provider that provides customers with access to one or more mobility services.

Network Edge
One or more locations within a network domain in close adjacency to the source of the data producer/consumer.

V2Cloud
Communication between a vehicle and applications or services hosted on a cloud.

Vehicle System
A system composed of a computing platform, applications, services and other components residing in the connected vehicle.

Abbreviations:

3GPP	The 3rd Generation Partnership Project
5G	5th Generation standard for broadband cellular networks
5GMS	5G Media Streaming
5GS	5G System
ACDC	Application-specific Congestion control for Data Communication
ADAS	Advanced Driver Assistance System
AF	Application Function
AMF	Access Management Function
ANDSF	Access Network Discovery and Selection Function
API	Application Programming Interface
APN	Access Point Name
BDT	Background Data Transfer
BSF	Bootstrapping Server Function
C-V2X	Cellular Vehicle-to-Everything
CAN	Controller Area Network
CUPS	Control User Plane Separation
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
DRB	Data Radio Bearer
DSRC	Dedicated Short Range Communication
ECU	Electronic Control Unit
eNB	Evolved Node B (LTE Base Station)

EPC	Evolved Packet Core
EPS	Evolved Packet System
EN-DC	E-UTRA-NR Dual Connectivity
FQDN	Fully Qualified Domain Name
GMA	Generic Multi-Access
gNB	Next Generation Node B (5G Base Station)
GPRS	General Packet Radio Services
GSMA	Global System for Mobile communications Association
GTP	GPRS Tunneling Protocol
HSS	Home Subscriber Server
HTTP	HyperText Transfer Protocol
IaaS	Infrastructure as a Service
IMEI	International Mobile Equipment Identity
IoT	Internet of Things
ISMP	Inter-System Mobility Policy
ISRP	Inter-System Routing Policy
IPSec	Internet Protocol Security
LTE	Long Term Evolution (the 4th Generation Mobile Communication Radio)
LWA	LTE WLAN Aggregation
MAC	Medium Access Control
M2M	Machine to Machine
MaaS	Mobility as a Service
MME	Mobility Management Entity
MNO	Mobile Network Operator
MSP	Mobility Service Provider
MPQUIC	Multi-Path QUIC
MPTCP	Multi-Path TCP
N3IWF	Non-3GPP Inter-Working Function
NAPT	Network Address Protocol Translation
NAT	Network Address Translation
NEF	Network Exposure Function
NF	Network Function
NR	New Radio (the 5th Generation Mobile Communication Radio)
OMA	Open Mobile Alliance
oneM2M	International standard organization for the Internet of Things
OS	Operating System
OTA	Over the Air
PaaS	Platform as a Service
PCC	Policy and Charging Control
PCF	Policy Control Function

PCRF	Policy and Charging Rules Function
PDN	Packet Data Network
PDU	Packet Data Unit
P-GW	Packet Gateway
P-GW-U	P-GW User Plane
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RAN	Radio Access Network
RAT	Radio Access Technologies
RRC	Radio Resource Control
RSD	Route Selection Descriptor
RTT	Round Trip Time
SD-WAN	Software-Defined Wide Area Network
S-GW	Serving Gateway
S-GW-U	S-GW User Plane
SCEF	Service Capability Exposure Function
SINR	Signal to Interference plus Noise Ratio
SIPTO	Selected IP Traffic Offload
SLA	Service Level Agreement
SMF	Session Management Function
SSC	Session and Service Continuity
SSID	Service Set Identifier
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UDR	User Data Repository
UDT	Unattended Data Traffic
UE	User Equipment
UL-TFT	Uplink Traffic Flow Template
UPF	User Plane Function
URSP	UE Route Selection Priority
USIM	Universal Subscriber Identification Module
UUID	Universally Unique Identifier
WLAN	Wireless Local Area Network
XMPP	Extensible Messaging and Presence Protocol

1 Introduction

Connected vehicles are expected to be a significant factor contributing to the growth of communications data volumes, with forecasts projecting every new vehicle produced being “connected” by 2025 [1]. Millions of cars are already connected using 4G cellular access, and cellular broadband IoT connectivity (4G/5G) is expected to grow significantly through 2024, as outlined in the Ericsson Mobility Report [2] and visualized in the Ericsson Mobility Visualizer [3]. According to the Cisco Annual Internet Report [4], connected vehicles that incorporate applications such as fleet management, in-vehicle entertainment, internet access, roadside assistance, vehicle diagnostics, navigation and advanced driver assistance services will be the fastest-growing industry segment with respect to machine-to-machine connections [4]. Furthermore, many emerging automotive services, such as intelligent driving assistance and Mobility as a Service (MaaS), work on the expectation that vehicles will be connected to cloud computing facilities. The AECC estimates that the related data traffic globally has the potential to exceed 10 exabytes per month by 2025, a volume 1,000 times larger than the present numbers, as described in an AECC white paper [5].

Stakeholders in this ecosystem, such as vehicle manufacturers, technology solution vendors, network operators, cloud infrastructure and service providers, must establish a practical platform architecture capable of supporting the variety of Vehicle-to-Cloud (V2Cloud) services, considering that the global distribution of vehicle fleets, vehicle data communications and the processing of that data at scale present a significant challenge to the currently deployed architectures.

The cellular network is one of the major access networks for connected vehicles, and many functions and mechanisms have been standardized in the 3rd Generation Partnership Project (3GPP). However, the present work within 3GPP has not fully addressed the challenge of automotive “big data,” and there is a high risk that future network deployments and business models will fail to support the emerging needs of connected vehicles. The cellular vehicle-to-everything (C-V2X) communication considered in 3GPP, for example, mainly covers latency-sensitive safety applications and does not address the forecast data volume growth between vehicles and the cloud. The 5G cellular system will provide improved functionality for both capacity and low latency, but the automotive industry will continue to use a mix of cellular access network technologies for the foreseeable future. In addition, increased data volume aggregated into data centers will cause network and processing congestion that degrades the user experience of connected vehicles.

The AECC believes that the current mobile communication network architectures and cloud computing deployments are not fully optimized to effectively handle emerging requirements of connected vehicles at a global scale. In response, the AECC proposes the use of a “Distributed Computing on Localized Networks” solution concept to solve these issues, with a proposal for a system architecture able to accommodate the predicted volumes of data traffic from connected vehicles. The concept focuses on three main aspects, which are localized networks, distributed computing and local data integration.

The system architecture provides a framework that supports the distribution of computation processes across a set of localized networks, shown in Figure 1.

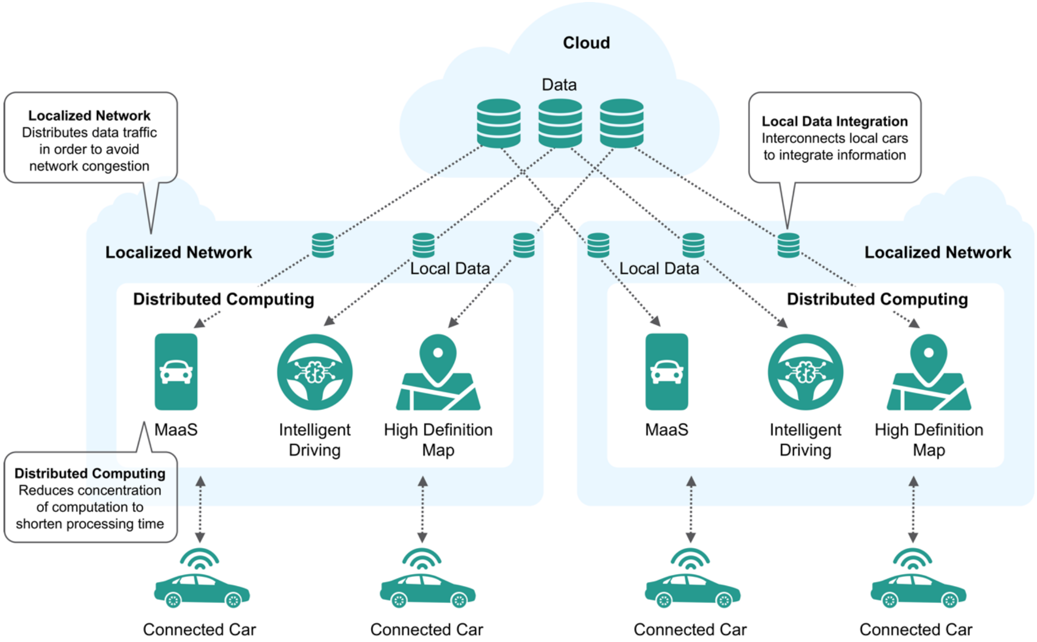

Figure 1. Three pillars of the AECC concept.

The AECC has identified a set of key issues, eight of which were prioritized for study, as shown in Figure 2.

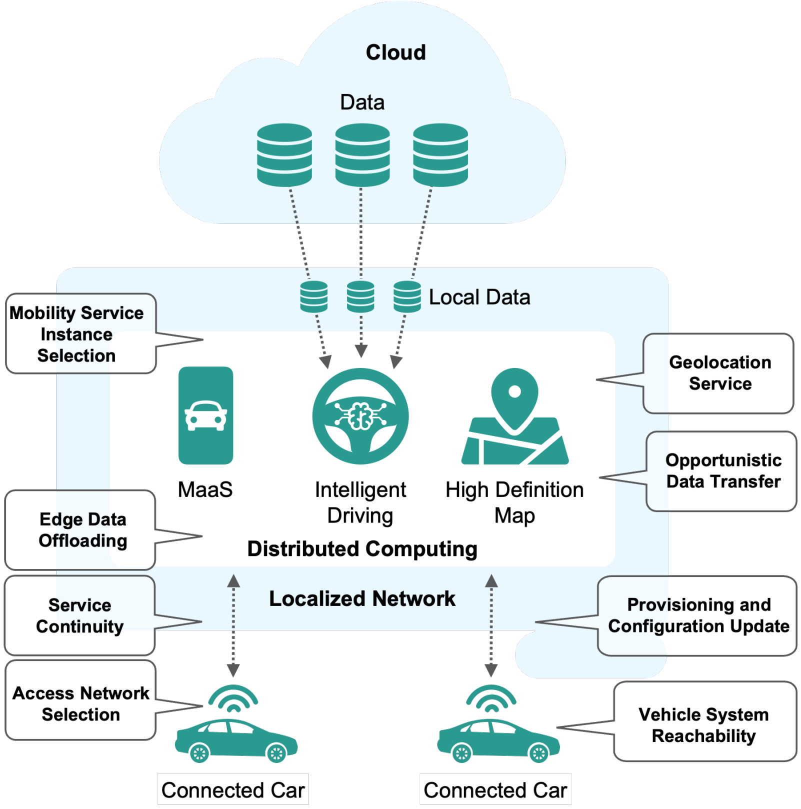

Figure 2. Eight key issues addressed by this technical report.

1. Edge Data Offloading – Cellular access is an important network access method for connected vehicles. Traditional cellular networks are designed to have a data gateway, through which all data to and from the connected vehicle will need to pass. The traditional approach creates a bottleneck and can prevent the efficient routing of data traffic into the distributed computing architecture. Discussion of this key issue describes how cellular networks can support the offloading of data to appropriate local computing infrastructure in an efficient and flexible manner.
2. Mobility Service Instance Selection – Mobility service instances are expected to make use of the capabilities of the distributed computing architecture. As application instances will be distributed across both edge and centralized computing infrastructure, this brings a new challenge for connected vehicles – how to determine which is the most appropriate mobility service instance to communicate with. Discussion of this key issue describes how connected vehicles will be able to select and use mobility service instances.
3. Vehicle System Reachability – Connected vehicles exhibit very different usage patterns and encounter more challenging connectivity conditions than most typical mobile devices, such as smartphones. With vehicle manufacturers responsible for globally distributed fleets of vehicle with heterogeneous network coverage, discussion of this key issue describes how connected vehicles can be contacted across heterogeneous network access methods.
4. Access Network Selection – To meet the requirements of mobility services, a connected vehicle is expected to use a mix of different access technologies and access networks. Determining which network to use may consider aspects such as network bandwidth, capacity, coverage and reliability. Discussion of this key issue introduces the ways in which a connected vehicle can select from two or more access networks according to service requirements and network capabilities.
5. Provisioning and Configuration Update – With mobility services needing to be provisioned in a flexible manner, adjusting to the changes within the numbers of locations of connected vehicles, it is essential to be able to direct how connected vehicles should use the available services. The required policy may also need to adapt, reflecting the changing services environment. Discussion of this key issue describes how configuration parameters and policies can be provisioned to connected vehicles.
6. Opportunistic Data Transfer – Bandwidth is a resource that needs to be carefully managed in cellular networks. Discussion of this key issue investigates how cellular network resources can be best leveraged providing data transfer for latency-tolerant mobility services used by connected vehicles while minimizing the impact to other users of the cellular network.
7. Service Continuity – In the AECC distributed computing architecture, a vehicle system may be served from different mobility services and IP anchors. Differing from edge data offloading and mobility service instance selection, this key issue only considers the case where handover of both mobility service instance and IP anchor takes place.
8. Geolocation Services – Given a set of thousands of geolocation events at any moment and millions of vehicles that change their locations each second and may also change their IP anchors, the network needs to calculate which vehicle needs to receive which notification using what network-provider-assigned IP address.

In order to address these issues, the AECC identified and evaluated a range of potential solutions and has provided recommendations on the different options.

2 System Overview

2.1 Architectural Requirements

The AECC proposes a system that will support the deployment and execution of mobility services using a distributed computing and networking architecture, including the vehicle system, networks (cellular network, WLAN, fixed access, IP-based, etc.) and center/edge servers. The primary goal of the architecture is to provide computing resources closer to where vehicle fleets are operating, enabling the processing of data to take place "in-region," instead of pulling data back to centralized cloud environments or centralized infrastructure operated by vehicle manufacturers, vehicle fleet operators and so on.

The AECC system architectural requirements related to the cellular network shall apply to both 4G and 5G systems. Also, they shall apply to non-standalone 5G deployments, where the core network is 4G and the 5G New Radio (NR) is used as radio access for data communication.

The vehicle system may connect to the AECC system using a WLAN when applicable. The WLAN may use Wi-Fi (IEEE802.11) as an access technology or 3GPP-based access technology (e.g., NR in millimeter wave carriers). The WLAN can connect to the cellular core network or to an internet service provider network.

The AECC system includes a distributed computing environment that will be used to support the various mobility services. Mobility services are executed within the vehicle system and computing facilities that may be composed of a center server and/or edge servers located outside the vehicle system. The investigation of distributed computing in the AECC is summarized in a separate white paper to be published in the future.

2.2 E2E Networking Reference Architecture for Hierarchical Data Traffic Distribution

Taking the intelligent driving service as an example, the AECC system introduces a hierarchical data processing architecture as shown in Figure 3, where the data will be processed not only in the vehicle system and the cloud but also at the edge servers located between the cloud and the vehicle system.

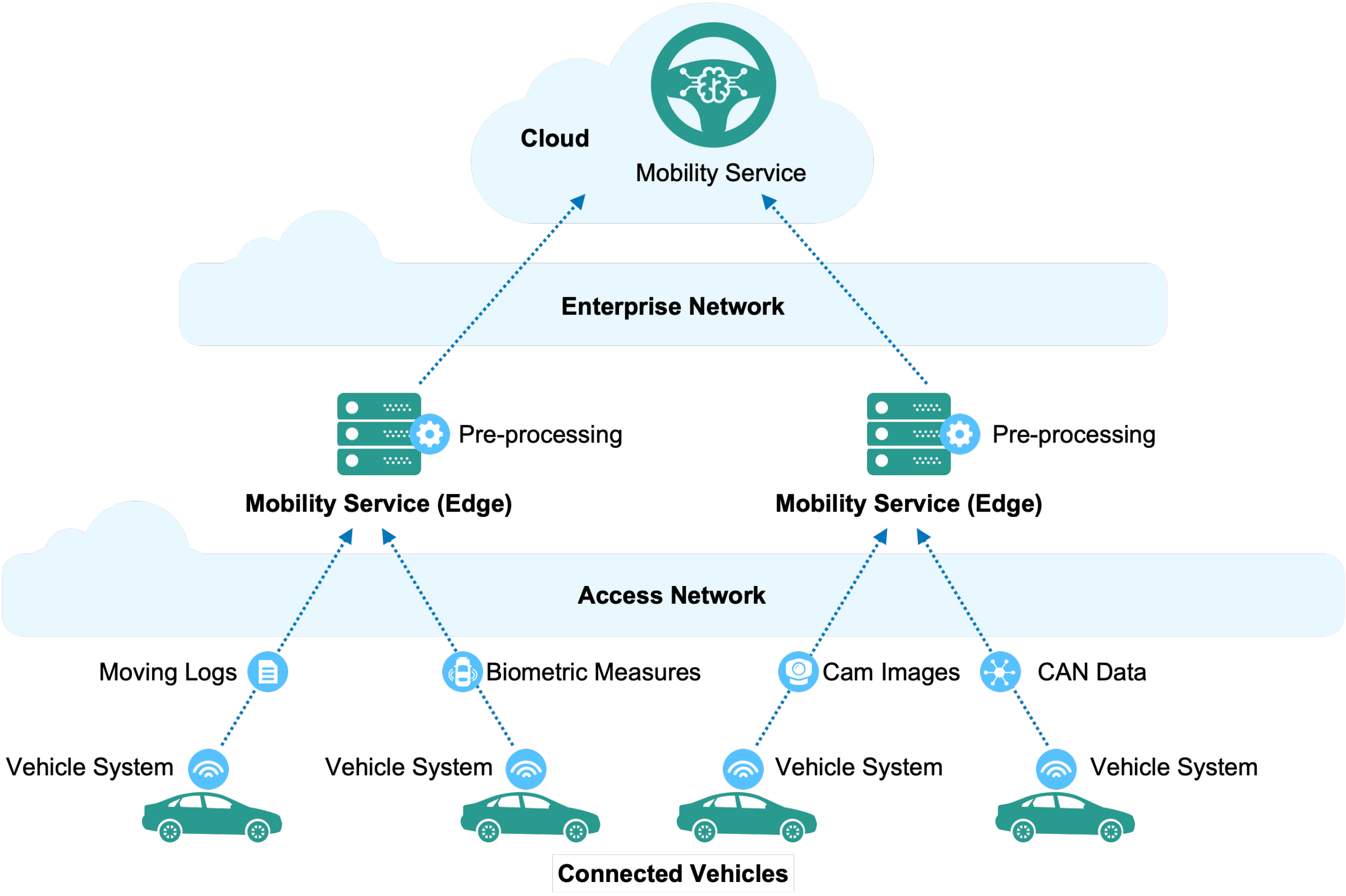

Figure 3. An example of the AECC system hierarchical data processing architecture.

2.2.1 Cellular Network Reference Models

In all cases of using cellular networks, the AECC system assumes the use of 3GPP reference models. This paper is written with the assumption that the reader is familiar with 3GPP reference models and associated terminology.

For the case where 4G is the only radio access used, the Evolved Packet System (EPS) architecture reference model is assumed [3].

For the case where non-standalone 5G NR is used as the radio access, the EPS architecture reference model according to 3GPP Release 15 and later (see [6]) with a non-standalone architecture option is assumed; i.e., 5G radio access is used in conjunction with an Evolved Packet Core (EPC). Figure 4 shows a non-roaming architecture as an example.

The 5G System (5GS) architecture reference model 3GPP 5GS Release 15 (see [7]) or later is assumed. This includes standalone NR deployment and multiple access dual connectivity options, with both NR and LTE radio connection to a 5G Core (5GC). Figure 5 shows a non-roaming architecture as an example.

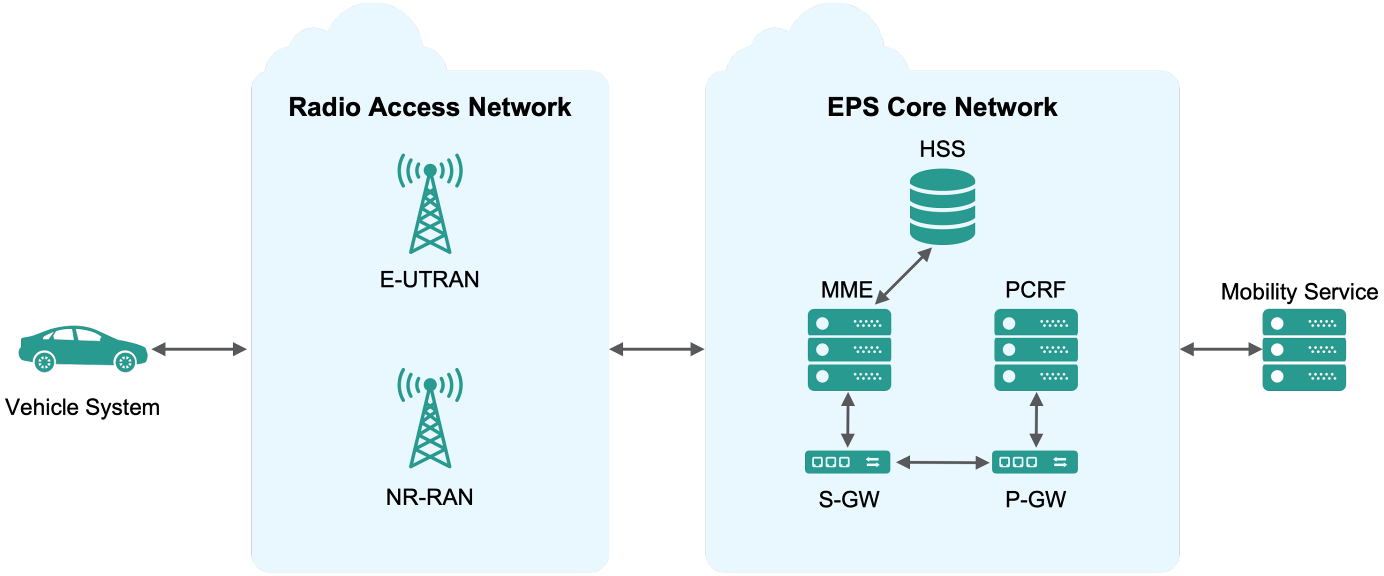

Figure 4. EPS architecture reference models with non-standalone NR radio access.

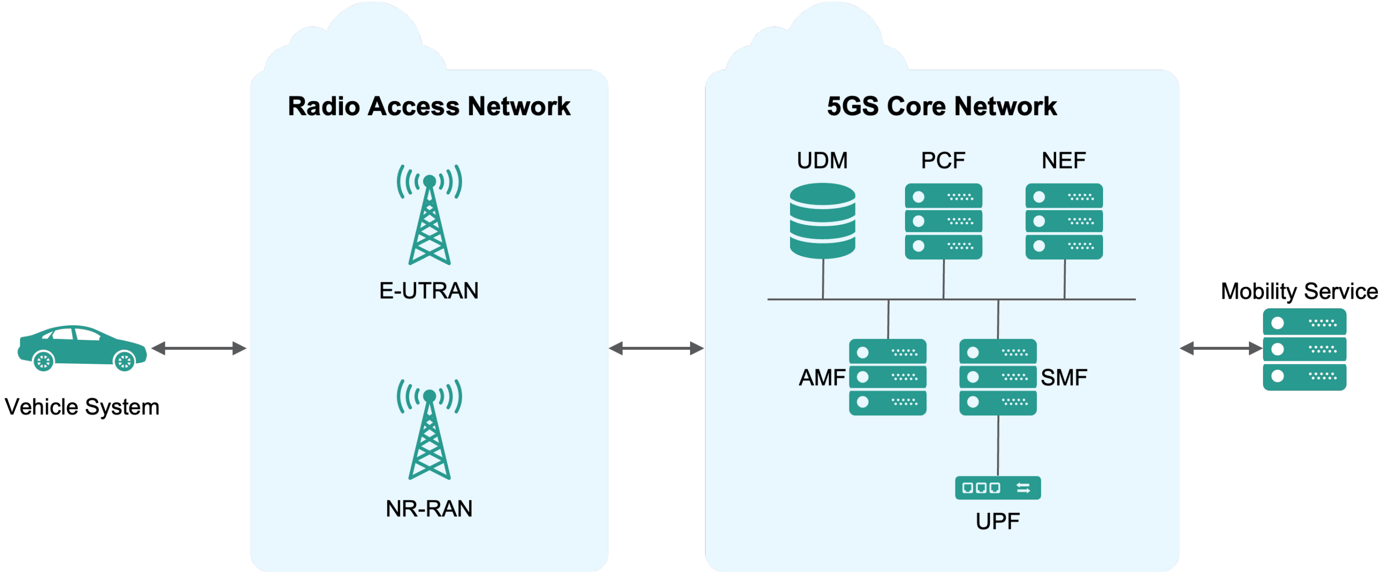

Figure 5. 5GS architecture reference models.

3 Key Issues and Solutions

3.1 Edge Data Offloading

3.1.1 Key Issue

In traditional cellular networks, a device connected to the network will have a fixed data offloading point in the network, through which all data to and from the device must pass. With connected vehicles, the movement may require frequent changes of the data offloading points and the serving mobility service instances.

The AECC system architecture envisages the use of distributed computing facilities. To use these facilities in an efficient and effective manner, cellular networks must also be able to change the selected data offloading point for a vehicle system as it moves, to provide an optimal communications path between the vehicle system and MSP edge servers. Furthermore, to optimize utilization of the cellular network’s edge infrastructure, only traffic designated to MSP edge servers should be offloaded at such edge data offloading points, while communication intended for other destinations, such as the MSP center server, should use more appropriate data offloading points, if available.

By using edge data offloading, cellular network operators will be able to take data communications traffic off the cellular network sooner, relieving potential congestion. In addition, edge data offloading is key to being able to direct data communications to local application instances, rather than to centralized instances.

In order to alleviate the impact of high-volume data transfers on cellular networks as identified in the AECC white paper [7], the cellular network (both EPS and 5GS), shall support data offloading to the designated MSP edge servers, as shown in Figure 8. The MSP edge servers are connected to the MSP center server via the enterprise network. In a cellular network, all traffic must enter and leave the network at specific data offloading points. According to the deployment of MSP edge server instances, these data offloading points shall be placed at appropriate locations in the cellular network to meet the service requirements on latency and capacity.

Note 1: the traffic flows of different services may selectively offload to different MSP edge servers to meet the various requirements of service use cases.

Note 2: the case of data offloading when using different access networks, such as WLANs, is discussed in Section 3.4, Access Network Selection.

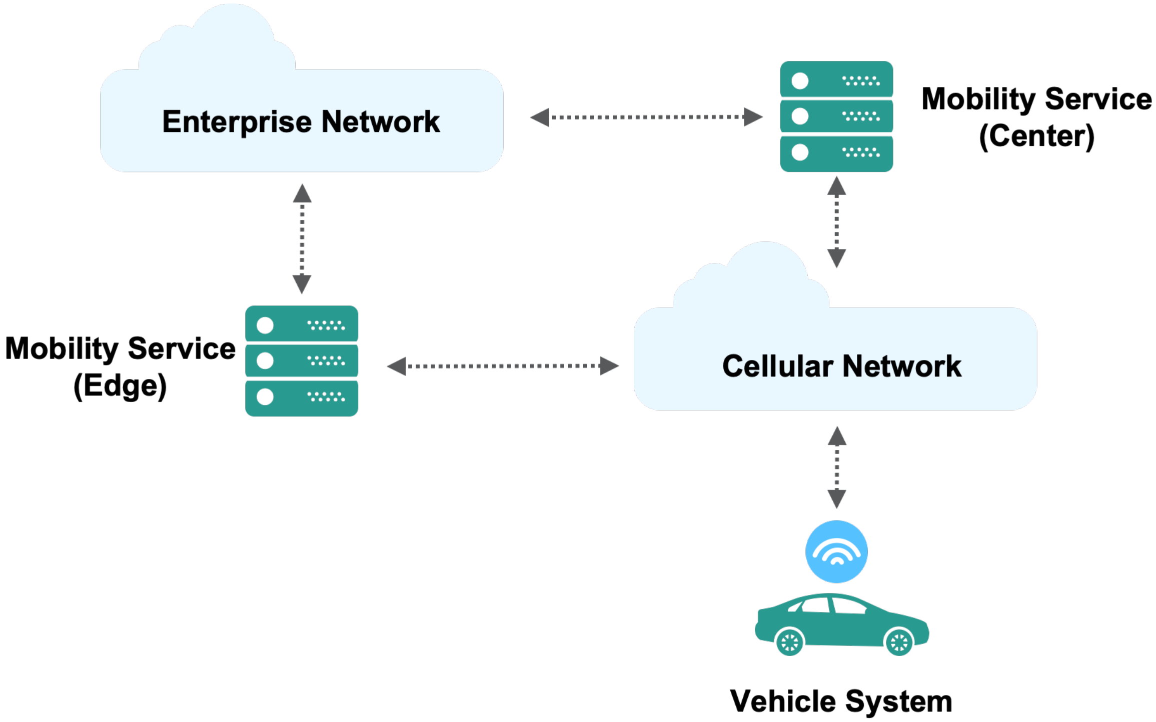

Figure 8. Connectivity between the vehicle system and mobility service instances can be provided over a cellular network, in which case appropriate data offloading points must be selected in the cellular network.

3.1.2 Potential Solutions

The following solutions are described for this key issue.

- □ Solution 1 – Data Offload with Single PDN Connection in EPS
- □ Solution 2 – Data Offload with Multiple PDN Connections in EPS
- □ Solution 3 – S1/N3 GTP Packet Filtering in EPS and 5GS
- □ Solution 3 – S1/N3 GTP Packet Filtering in EPS and 5GS
- □ Solution 4 – Data Offload with Single PDU Session in 5GS
- □ Solution 5 – Data Offload with Multiple PDU Sessions in 5GS
- □ Solution 6 – Uplink Classifier
- □ Solution 7 – IPv6 Multi-Homing

The first section below introduces the general problem statement and the applicability of different solutions to different situations. Afterward, all solutions are defined one by one.

3.1.2.1 Solutions Overview

Edge Data Offloading in the Evolved Packet System

In 4G, or more specifically in the Evolved Packet System (EPS), PDN-Gateways (P-GWs) act as data offloading points of the cellular network. The Control and User Plane Separation (CUPS) feature provides more flexibility when it comes to data offloading and splits the P-GW functionality into a control plane entity (P-GW-C) and a user plane entity (P-GW-U). Once a PDN connection from user equipment (the UE being the communications module within the vehicle system) to such a P-GW-U is established, it cannot be re-anchored to a different P-GW. Provided the PDN connection persists, all traffic will be offloaded to the initial P-GW-U (see Figure 9).

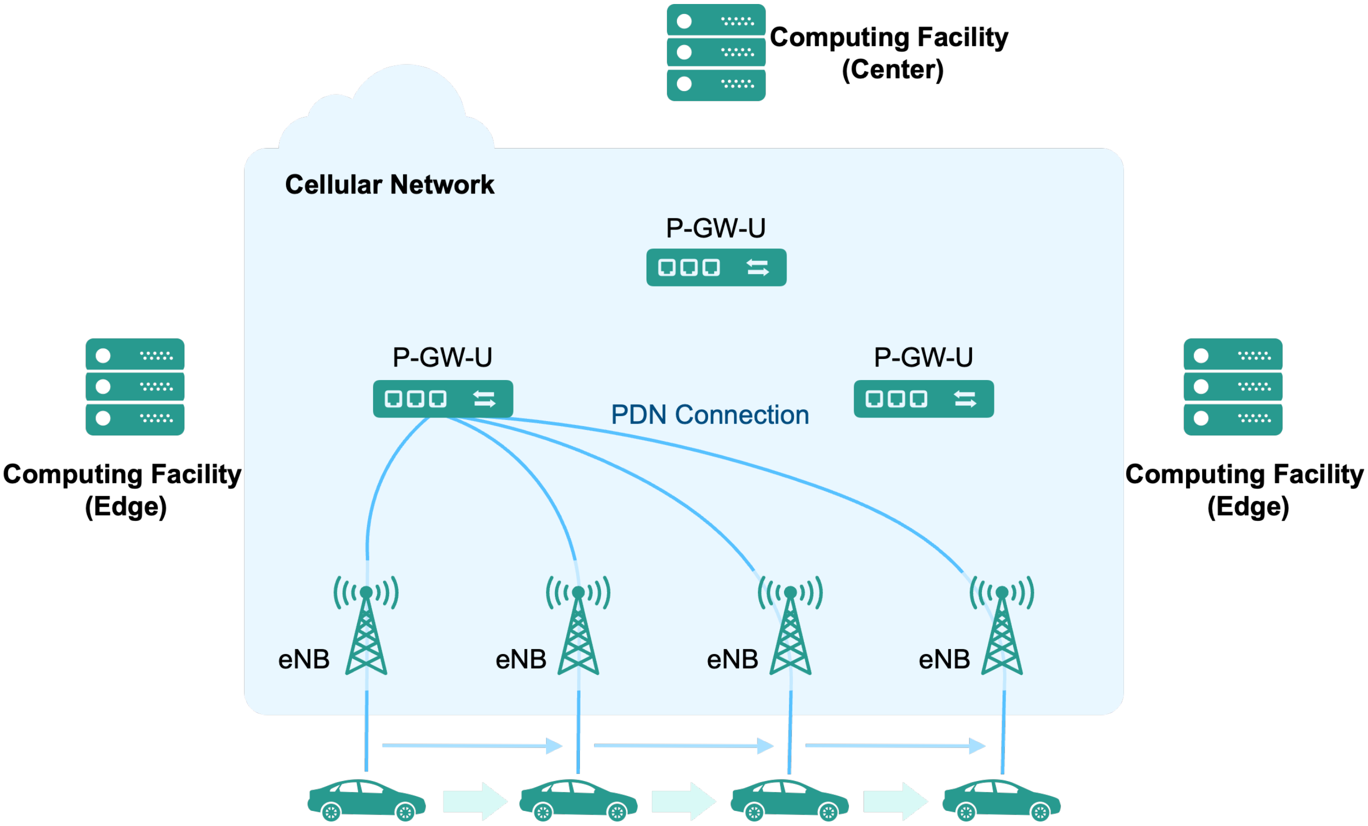
Computing Facility (Center)
Cellular Network
P-GW-U
P-GW-U
P-GW-U
PDN Connection
Computing Facility (Edge)
Computing Facility (Edge)
eNB
eNB
eNB
eNB

Figure 9. Vehicle moving with a persisting PDN connection in EPS.

For re-initiating the P-GW-U selection procedure, the PDN connection must be reactivated or a new PDN connection must be established. The latter is done during a re-attachment procedure, which automatically happens when connectivity is re-established after it was lost for some time, such as due to large radio coverage white spots. During the attachment procedure, the MME will then select an appropriate P-GW-U for the PDN connection, based on the operator's configuration, such as tracking area.

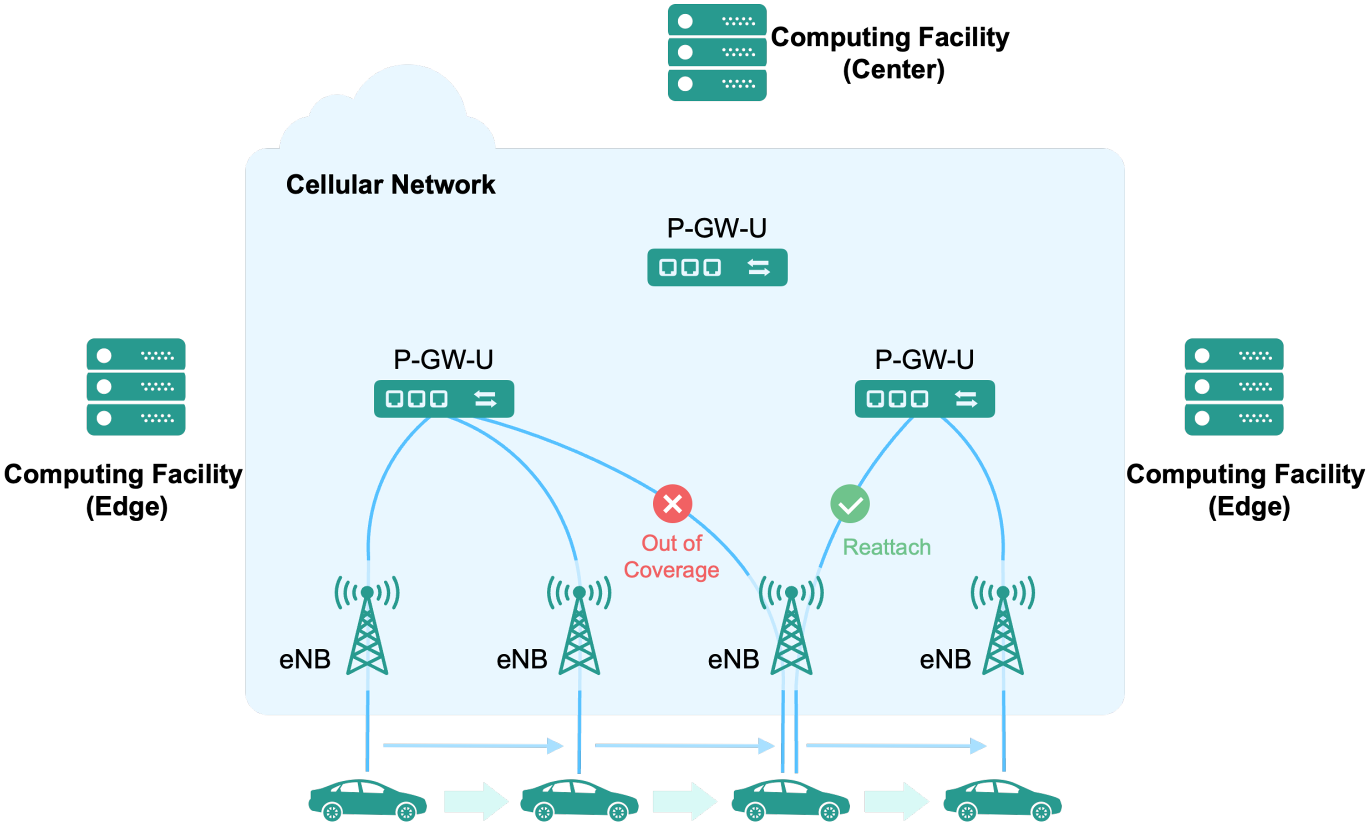
Computing Facility (Center)
Cellular Network
P-GW-U
Computing Facility (Edge)
P-GW-U
P-GW-U
Computing Facility (Edge)
Out of Coverage
Reattach
eNB
eNB
eNB
eNB

Figure 10. Vehicle moving with changing PDN connections in EPS.

When running into a large coverage white spot, the PDN connection is terminated and a new PDN connection is created when returning to coverage. This ensures proper selection of an appropriate P-GW-U. If a cellular network has several white spots between areas that should be served by different P-GW-Us,[1] movement between areas served by different P-GW-Us might not be an issue, as the P-GW-U is automatically reselected on a regular basis, if the network is configured accordingly. However, with many P-GWs deployed and more continuous coverage, a PDN connection reactivation must be triggered by the system. For this purpose, EPS provides an optional feature (“Selective IP Traffic Offload [SIPTO] above RAN”) where the MME requests a PDN connection deactivation with a subsequent reactivation to the UE (see Figure 11).

[1] Note: this is a typical case for spotty coverage in rural areas with a low degree of P-GW-U distribution.

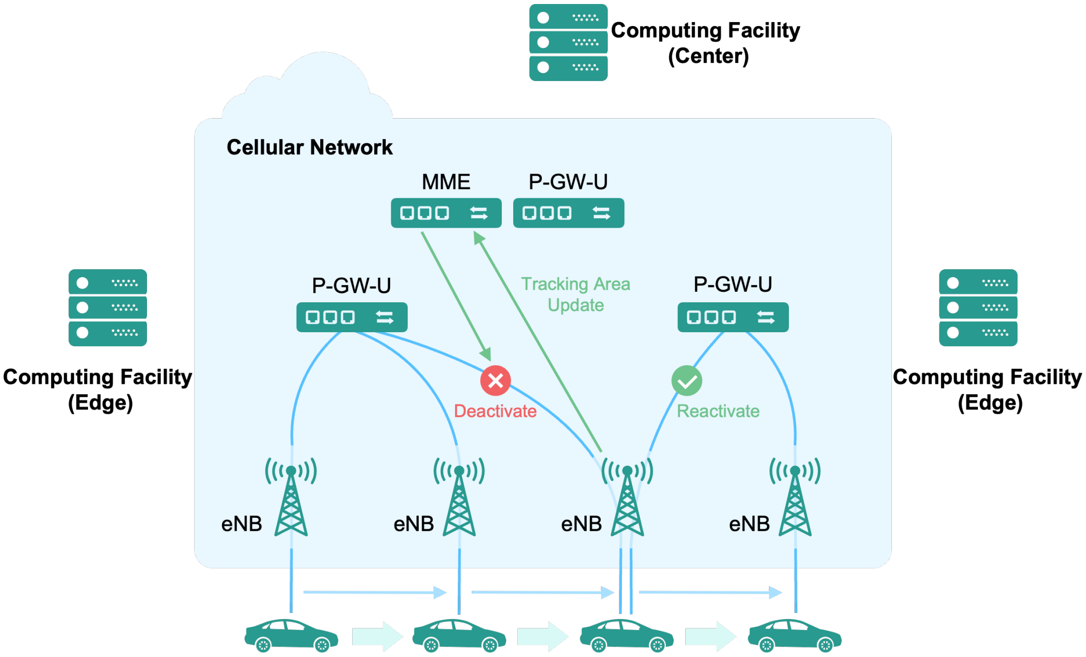

Figure 11. Reactivation of the PDN connection with SIPTO.

When using only a single PDN connection per UE, data exchanges between a vehicle system and an MSP center site will go via the current edge P-GW-U (see Figure 12, right car). This may require over-dimensioning the edge P-GW-U if the share of data is relevant and geographically varying. In that case, the load on the edge P-GW-Us can be reduced by using a second PDN connection on the same UE, with a different APN that is configured for anchoring at a central P-GW-U (see Figure 12, left car). In that case, two IP interfaces must be managed on the vehicle system, one per APN.

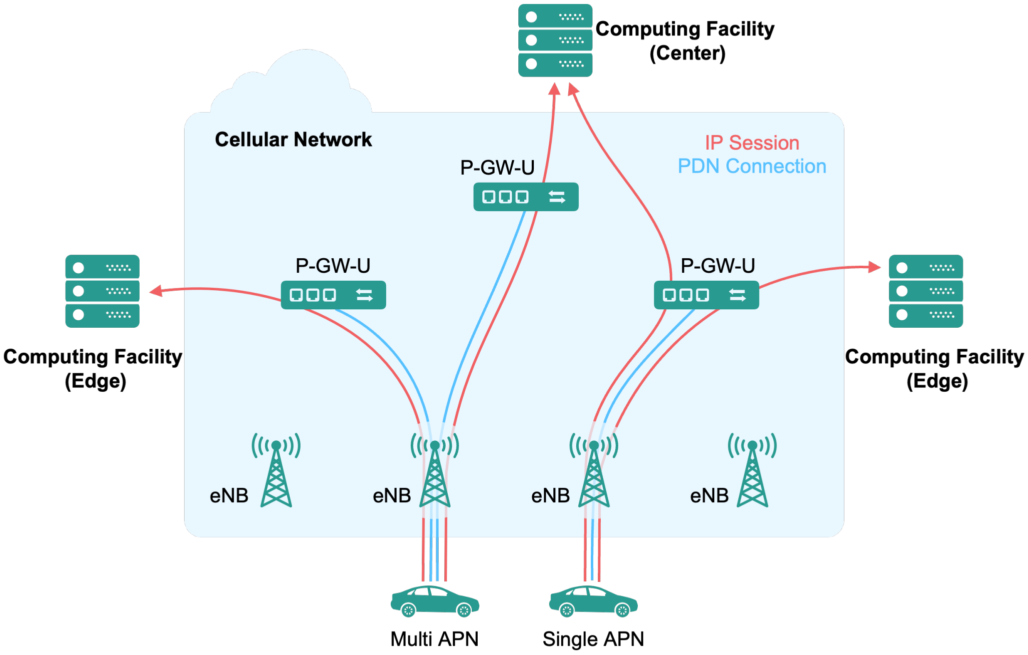

Figure 12. Offloading edge GWs with a second PDN connection.

Edge Data Offloading in the 5G System

The 5G System (5GS) has the same paradigm of an existing central anchor point (the PDU session anchor in the User Plane Function [UPF]). This requires similar solutions with a single PDU session that offloads at the edge. However, compared to the EPS, the 5GS supports multiple Session and Service Continuity (SSC) modes that control behavior during mobility between anchor points (UPFs). Depending on the SSC mode, a PDU session will persist until running out of coverage (SSC mode 1), a PDU session may be released and immediately re-established to a new UPF (SSC mode 2, SIPTO-like behavior) or a new PDU session may be established before releasing the old PDU session (SSC mode 3). Just as in EPS, an additional PDU session, anchored at a central UPF, can be used to reduce the load on edge UPFs.

While the features described above enable a vehicle system to always have connectivity via appropriate data offloading points, ongoing sessions are interrupted during re-attachment procedures, meaning that data communication must be re-established, resulting in a temporary interruption in service. The 5GS also offers mechanisms to maintain connectivity while re-anchoring the PDU session at a different UPF. For example, an uplink classifier (ULCL) policy can be provisioned in a UPF to offload the selected traffic to an edge server, and the SMF can dynamically insert and remove an uplink classifier into the data path of the PDU session (see Figure 13). Typically, IP 5-tuples are used in such policies to decide which packets to offload to which edge site.

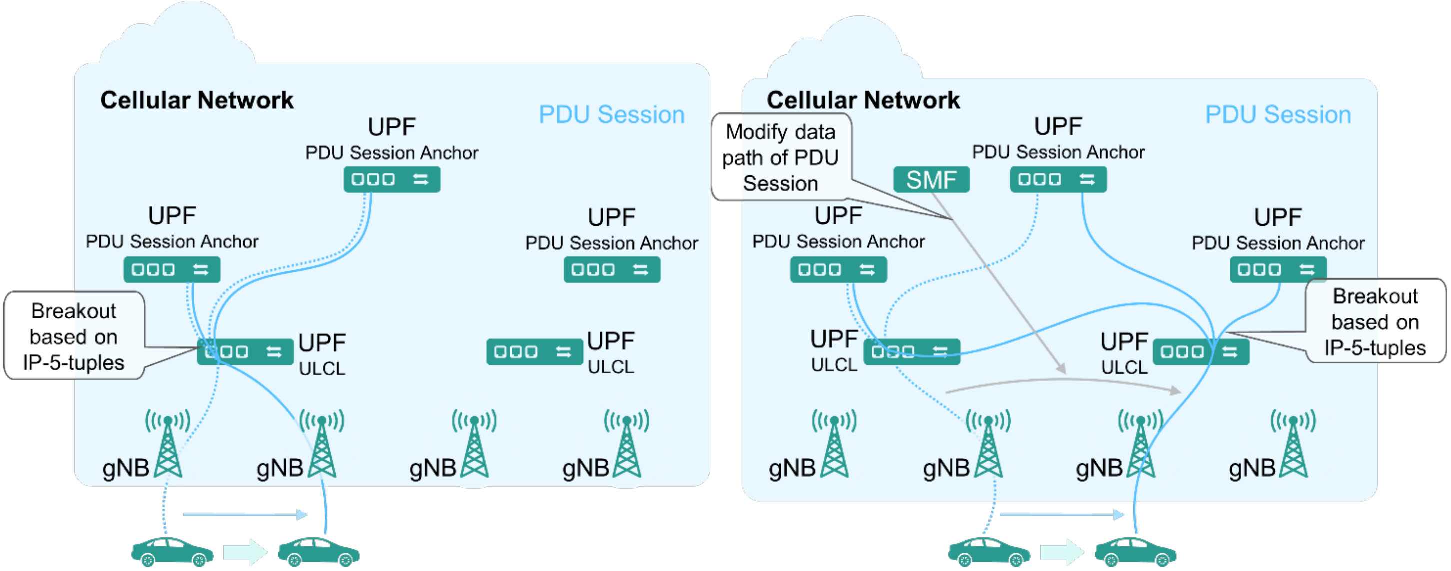

Figure 13. Selectively routing packets with uplink classifiers.

In this approach, while there are multiple PDU session anchors, there is only one IP anchor. Even when moving the PDU session to a different PDU session anchor (UPF), the IP session is maintained while traffic to the old uplink classifier UPF is tunneled. This communication link is deactivated when not used anymore, such as when using timeout procedures (see Figure 14). For downlink traffic, the vehicle system is reachable using the same IP via all UPFs, which must be taken into account when considering IP routing and path selection between the mobility service instances, the intermediate IP network(s) and the UPFs themselves. As a consequence, this solution comes with significant complexity in the network design, where the optimal downlink routing decision depends on the location of the UE only known by the mobile network.

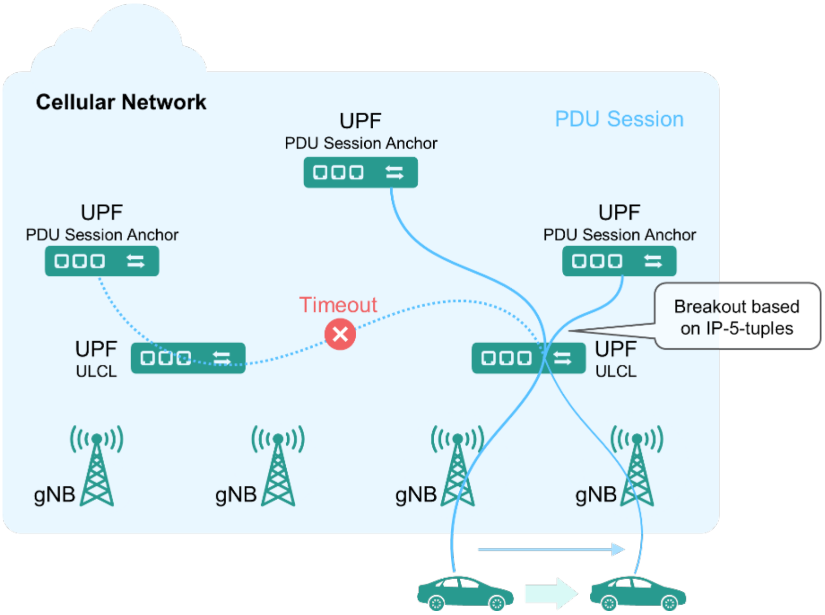

Figure 14. Terminating an unused link with timeout procedures.

3.1.2.2 Solution 1 – Data Offload with a Single PDN Connection in EPS

In EPS, SIPTO above RAN enables dynamic reselection of GWs (S-/P-GW), and selection of GW-Us in the case of Control and User Plane Separation (CUPS). Those GWs or GW-Us are geographically/topologically close to the UE. The selection mechanism can consider UE location in the network (tracking area), APN or other parameters.

The UE does not need to know whether the PDN connection corresponds to the MSP edge or center server. One AECC system-dedicated APN can be provisioned to the UE for all AECC system traffic flows – including traffic with and without offloading. When the UE uses this AECC system APN, it does not need to know whether or not traffic on this APN will be offloaded. The network (MME) will choose appropriate GWs for AECC APN traffic purely based on the MNO's configuration; e.g., based on defined groups of cells (tracking area). The GW (S-/P-GW or S-/P-GW-U) selection will be based on information such as SIPTO permission information per subscription per APN, UE location information and so on. The MME can also decide to move a PDN connection from one GW to another (e.g., from a GW serving the MSP edge server to a GW serving the MSP center server) for an AECC system APN if needed.

The different types of AECC system traffic flow will be offloaded to the applications on the MSP edge server or pass through to the MSP center server as shown in Figure 15.

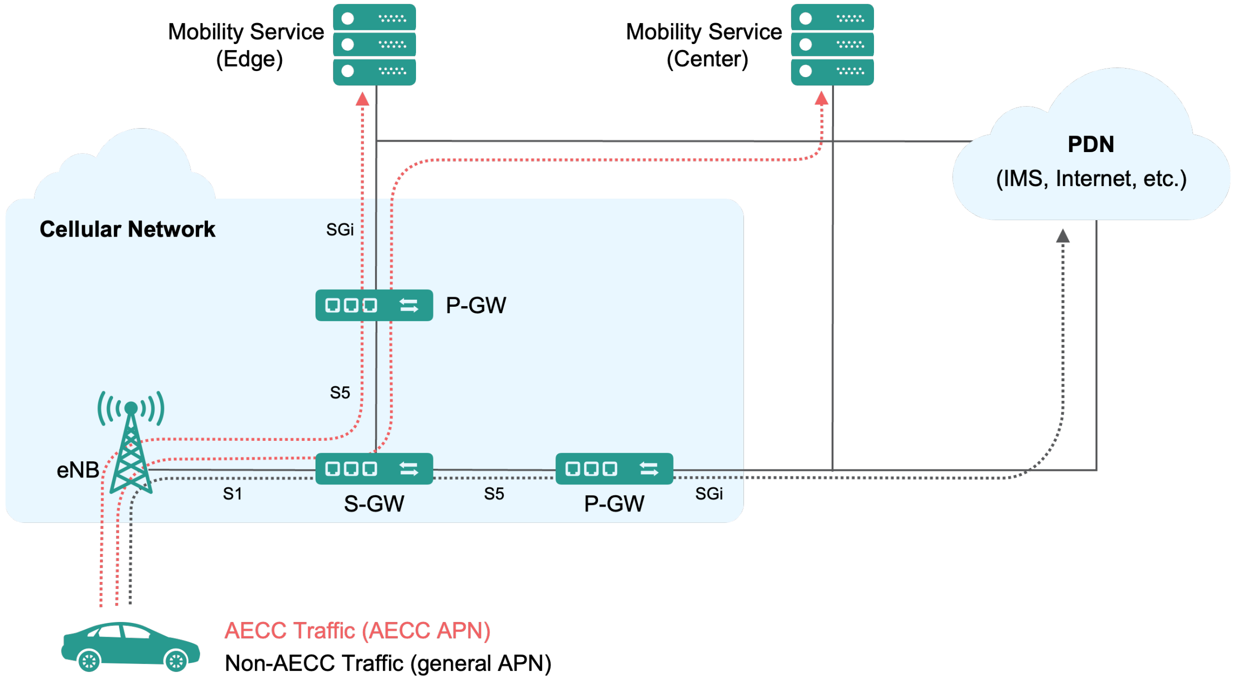

Figure 15. Data offload with a single PDN connection for AECC system-related traffic (red) and other traffic (black) in EPS.

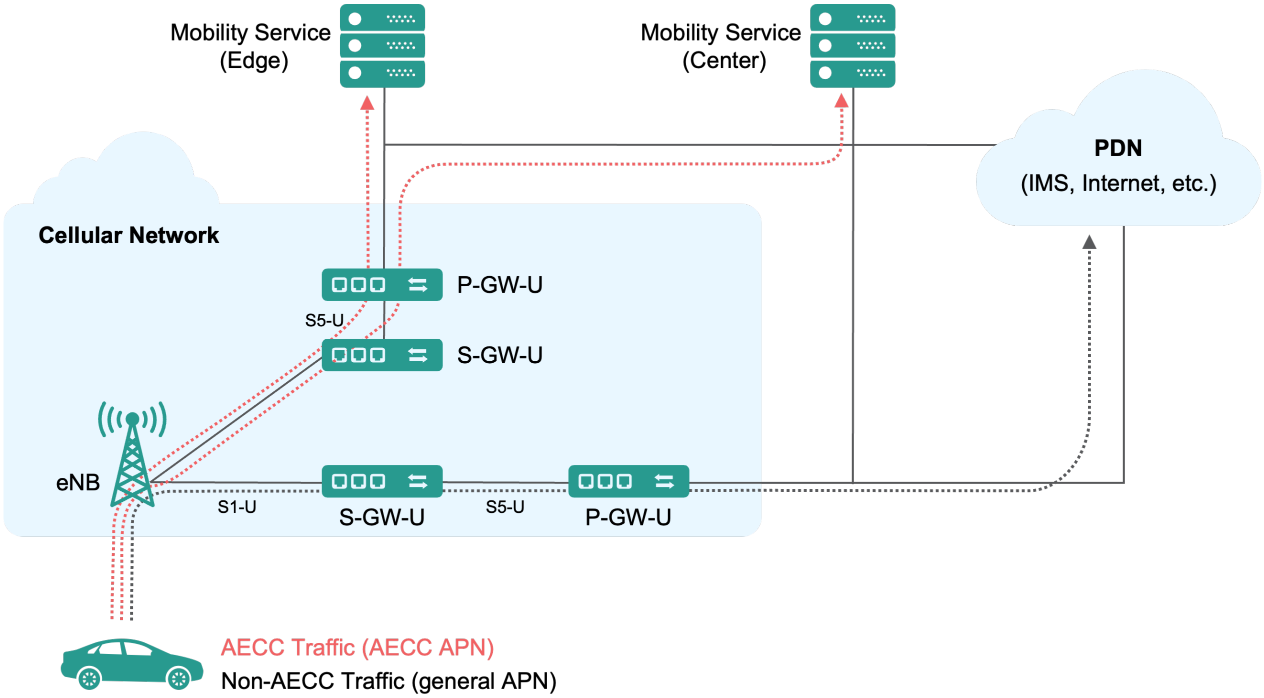

Figure 16. Data offload with a single PDN connection for AECC system-related traffic in EPS in the case where CUPS is deployed.

3.1.2.3 Solution 2 – Data Offload with Multiple PDN Connections in EPS

SIPTO above RAN enables the cellular network to offload data traffic sent through a PDN connection for an APN configured for SIPTO to a designated MSP edge server via a selection of P-GW (or P-GW-U in the case of CUPS). It is a function introduced and standardized since 3GPP LTE Release 10. SIPTO requires a dedicated APN for offloading the selected data traffic to the edge server, so it needs to support multiple APNs for different PDN connections at the UE side to achieve a selective data offload. Consequently, the vehicle system needs to implement support for multiple outbound IP interfaces (each corresponding to a different PDN connection), and corresponding routing functionality (which includes static routing or “hardwiring”) in order to place the IP traffic onto the correct interface.

Due to movement of the UE in the network, the serving MME may need to redirect a PDN connection to a different P-GW that is more appropriate for the location of the UE, based on the tracking area of the vehicle system. In this case, this solution cannot maintain session continuity while changing to the new P-GW.

The GW selection is configured by the MNO through tracking area configuration and mapping to GWs; that is, no direct control to external entities is provided in the current solution. For selecting which PDN connection to use for a packet in the uplink, one can either push the selection process to the application layer in the UE or use UL-TFTs for doing an automatic mapping based on IP 5-tuples. For the downlink, one can select a PDN connection by using the respective IP address assigned by the corresponding GW.

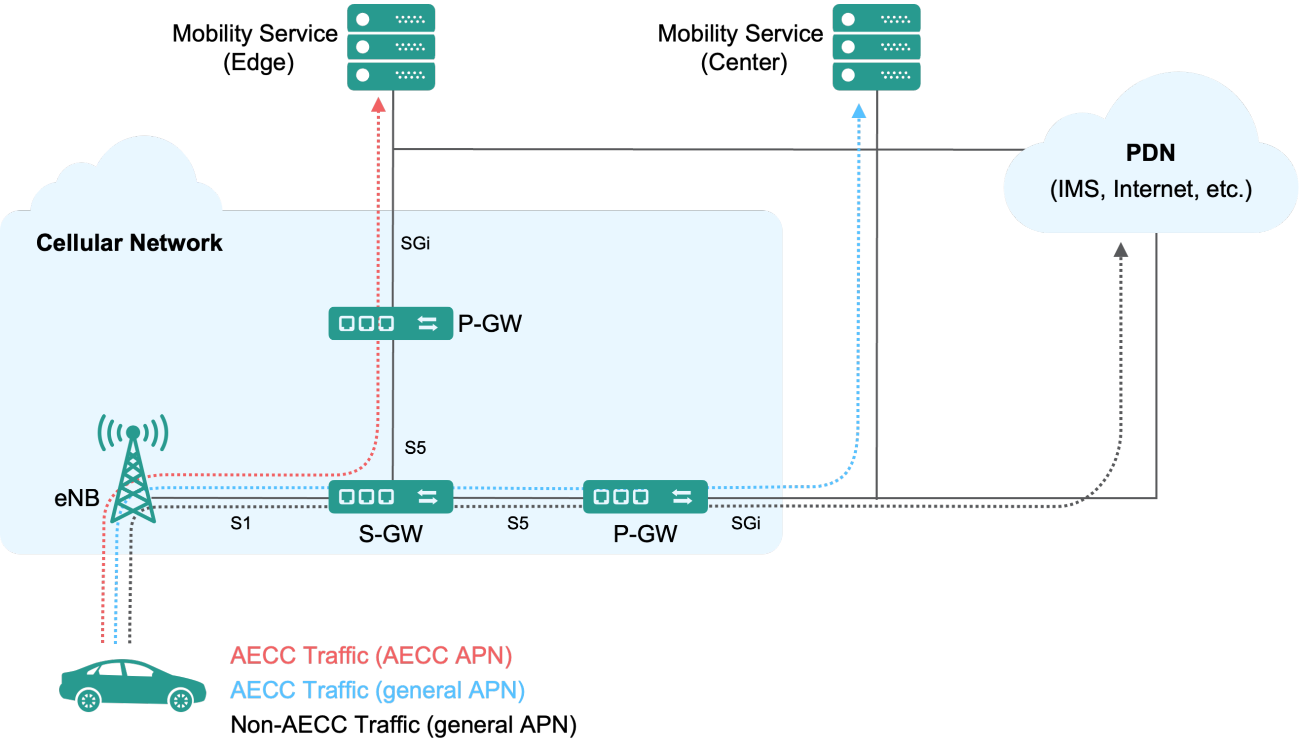

Figure 17. Data offload with multiple PDN connections for AECC system-related traffic (red and blue) and other traffic (black) in EPS.

3.1.2.4 Solution 3 – S1/N3 GTP Packet Filtering in EPS and 5GS

Edge data offloading can be conducted based on S1 GTP packet filtering mechanisms. The whole solution should include:

- Packet filtering, also called a traffic filter policy
- Packet filtering management

To support edge offloading in the cellular network, a traffic filter solution to handle the traffic offloading of the data plane can be used. For the packet filtering policy, a set of traffic filter policies composed of traffic rules and traffic filters corresponding to the rules is shown in Figure 18.

For example, such a TrafficFilter policy could use IP header information (IP address, port, L4 protocol) or consider L4/L7 parameters when feasible. If the edge offloading resides on the S1 interface inside the cellular network, it can also support filtering based on GTP tunnel information such as GTP-U TEID and so on. The policy also includes actions such as forward, drop, passthrough and duplicate.

For the packet filter policy (TrafficFilter policy) management, the system uses an architecture solution to address policy management for edge offloading inside 4G or 5G cellular networks. The AECC system can generate traffic rules in order to exert influence over the data plane within the packet filter in the cellular network.

The solution for the packet filter is shown in Figure 18, where packet filtering must be implemented on a device on the S1 interface between eNBs/gNBs and the core network.

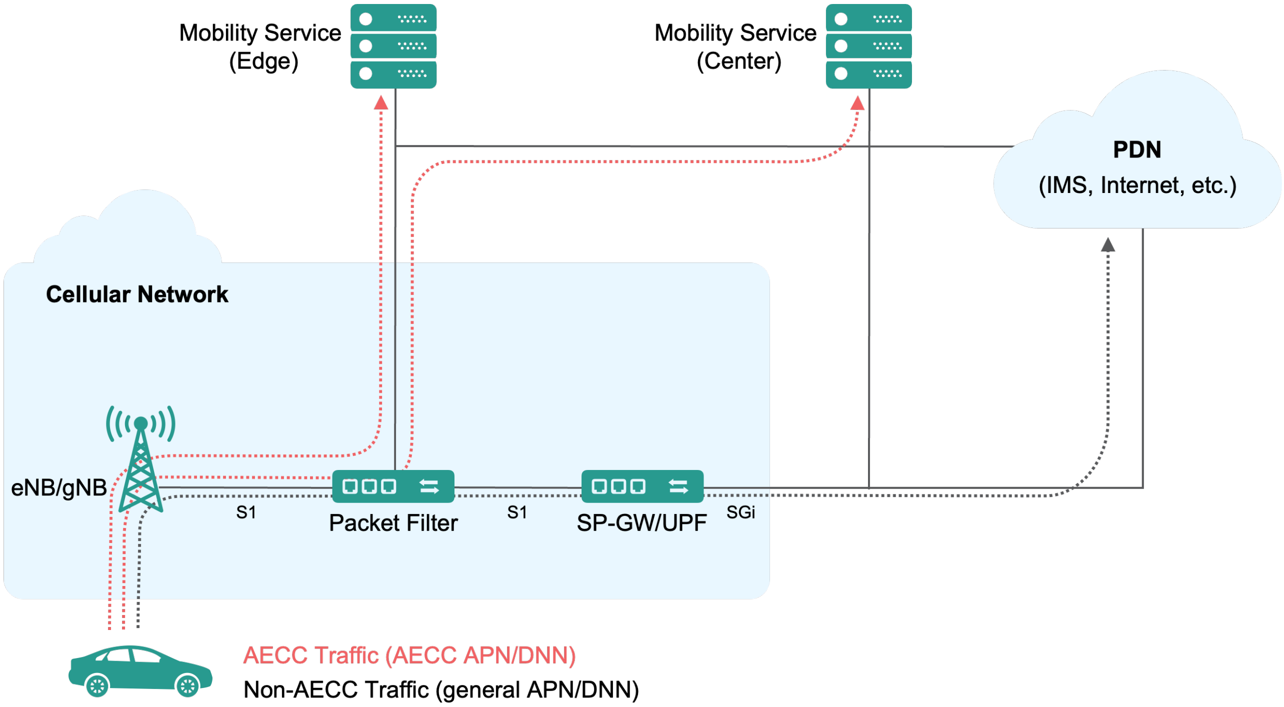

Figure 18. GTP packet filter implementing traffic filter policy.

In Figure 18, a packet filter policy based on port numbers, for example, can be used to apply different edge data offloading behaviors.

The key issue is solved in this case by intercepting S1-U (in EPS) or N3 (in 5GS) traffic and communicating directly with the mobility service instances from the intercepting entity.

This solution is applicable to both EPS and 5GS.

3.1.2.5 Solution 4 – Data Offload with a Single PDU Session in a 5GS

In a 5GS [7], the SMF is in charge of selecting or reselecting a UPF for a PDU session and can consider a number of parameters for the selection process. Among these is the tracking area identifier, which allows for a cell-specific UPF selection. In this approach, the selection process is used to offload all traffic to an edge UPF. Likewise, downlink traffic would always pass through this UPF. Different traffic flows might still be offloaded to different mobility service instances but would use the same breakout/UPF.

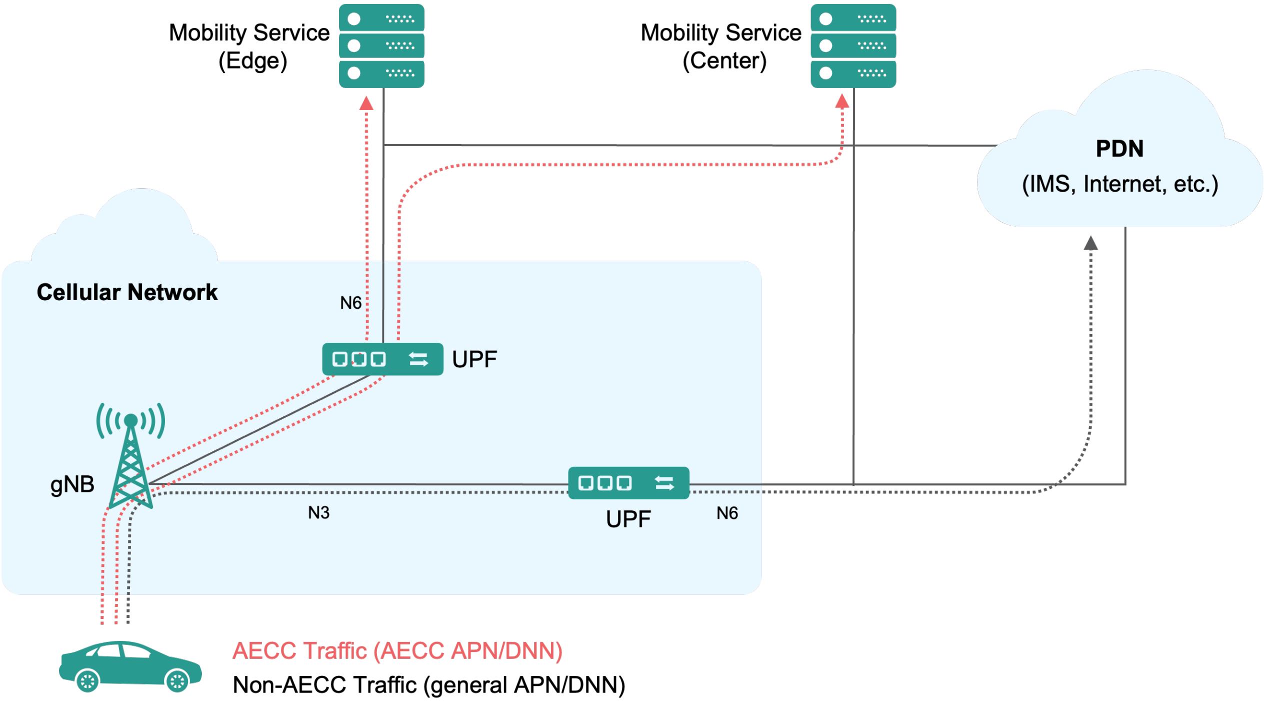
Mobility Service (Edge)
Mobility Service (Center)
PDN
(IMS, Internet, etc.)
Cellular Network
N6
UPF
gNB
N3
UPF
N6
AECC Traffic (AECC APN/DNN)
Non-AECC Traffic (general APN/DNN)

Figure 19. Data offload with a single PDU session for AECC system-related traffic (red) and other traffic (black) in a 5GS.

3.1.2.6 Solution 5 – Data Offload with Multiple PDU Sessions in a 5GS

The UE creates multiple PDU sessions using the same procedure as for the edge breakout with a single PDU session. While one PDU session offloads traffic to edge UPFs, the other PDU session uses a central UPF. For downlink traffic, the two external IP addresses of the respective PDU sessions are used to select the corresponding UPF. Again, the vehicle system needs to implement support for multiple IP interfaces (each corresponding to a different PDU session) and corresponding routing functionality.

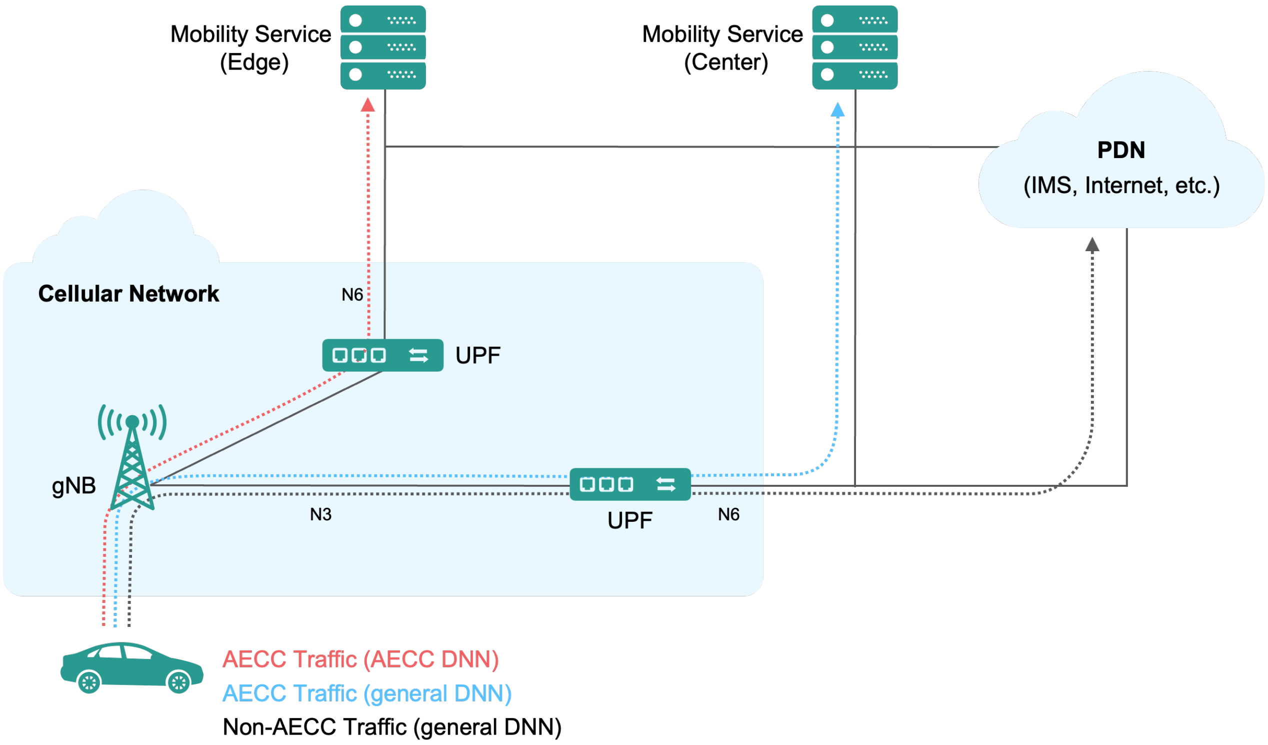

Figure 20. Data offload with multiple PDU sessions for AECC system-related traffic (red and blue) and other traffic (black) in a 5GS.

3.1.2.7 Solution 6 – Uplink Classifier

In a 5GS, an uplink classifier policy can be provisioned in a UPF to offload the selected traffic to an edge server. The insertion and removal of an uplink classifier policy is controlled by the SMF. The SMF may include multiple UPFs with uplink classifier policies in the traffic data path and may modify this UPF chain dynamically. This solution is only supported in a 5GS, and IP 5-tuples are used as traffic filters in the UPFs that, when matched, trigger local offload of the respective traffic.

In this approach, while there are multiple PDU session anchors, there is only one IP anchor; that is, the IP address of the UE is assigned by only one UPF and preserved during the lifetime of the PDU session. Even when the PDU session anchor changes (e.g., due to movement of the vehicle system), the IP session is maintained, while traffic to the old uplink classifier UPF is tunneled. For downlink traffic, the vehicle system is reachable using the same IP via all UPFs, which must be considered in the IP configuration of the mobility service instances and the corresponding IP network(s).

To forward data to the appropriate PDU session anchors, uplink classifiers must be configured accordingly, based on information on IP subnets and location of mobility service instances, in order to know which PDU session anchor is most appropriate for the UE in a given cell, and the IP subnet with which it communicates. This information can be configured and updated manually, or it can be dynamically exposed to an MNO by an MSP. Typically, defining how such information is exchanged is described as part of an SLA.

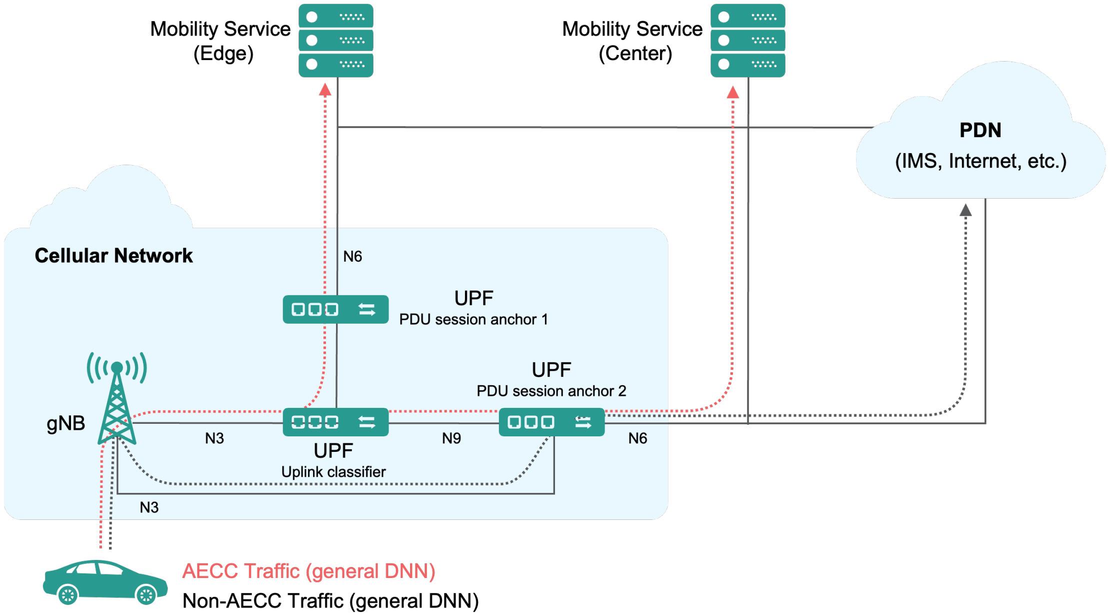

Figure 21. Data offload with a single PDU session for AECC system-related traffic (red) and other traffic (black) in a 5GS, using uplink classifiers to selectively offload traffic based on traffic filters (usually IP 5-tuples).

3.1.2.8 Solution 7 – IPv6 Multi-homing

In a 5GS, the PDU session from a vehicle system may be associated with multiple IPv6 prefixes. Selected traffic can be offloaded to the designated edge server as configured by the SMF, using a specific IPv6 prefix. In the traffic data path, the common UPF acts as a branching point, where the uplink traffic is split to different destinations and downlink traffic is merged to the vehicle system. The UE selects the source IPv6 prefix according to rules preconfigured in the UE or received from the network.

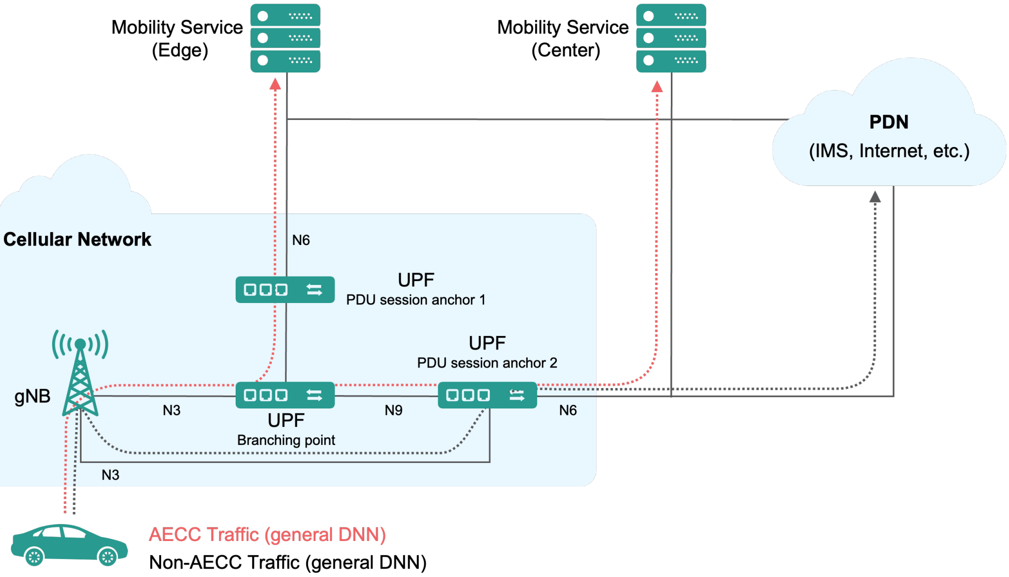

Figure 22. Data offload with a single PDU session for AECC system-related traffic (red) and other traffic (black) in a 5GS, using different IPv6 prefixes to selectively offload traffic.

3.1.3 Conclusions

The solutions recommended by the AECC are Solution 5 (Data Offload with Multiple PDU Sessions in 5GS) for 5GS, and Solution 2 (Data Offload with Multiple PDN Connections in EPS) for EPS. Both have very similar behavior and capabilities and are supported by default in any mobile network of the respective generation (4G and 5G). Many modern vehicle systems already support handling multiple parallel sessions with different APNs/DNNs, and can support these solutions with limited overhead.

As a fallback for a low-complexity vehicle system design, Solution 4 (Data Offload with a Single PDU Session in 5GS) and Solution 1 (Data Offload with a Single PDN Connection in EPS) can be used, if a vehicle system is not capable or not required to manage multiple sessions in parallel. These solutions have very low complexity and only require a subset of features compared to solutions 5 and 2; that is, they can easily co-exist with the recommended solutions in the same network, and even use a similar software design in the vehicle system.

In the long term, with 5G being widely available, supported and mature, Solution 6 (Uplink Classifier) is an interesting option, as it leaves IP session termination during anchor point changes to the applications, maintains IP connectivity via the central breakout point irrespective of the current edge breakout configuration and avoids complexity in the vehicle system. It also allows for the deployment of edge breakout capabilities with reduced capacity by comparison to solutions where all data must pass through the particular edge breakout function. Consequently, it can be used as an optimization where it is available, if the required integration with the AECC system is done and mobile network operators support it in their networks.

3.2 Mobility Service Instance Selection

3.2.1 Key Issue

The AECC system is expected to support the execution of software applications that will be used by vehicle systems of different types and from differing manufacturers. The working assumption is that applications will be delivered utilizing IPv4 or IPv6-based communications protocols and that in keeping with today's modern cloud deployment platforms, a dynamic mechanism will be required that will be able to inform the vehicle system of the resources that are available to it and then direct the vehicle system's application software to use the most appropriate application server instance. The selection function, hereafter referred to as the mobility service instance selection service, forms part of the overall set of services provided by the AECC system, enabling the exchange of data between applications executing in the vehicle systems and mobility service instances.

The working assumption adopted by the AECC is that where there are multiple concurrent applications in use within a vehicle system, the vehicle system may connect to multiple mobility service instances as shown in Figure 23, since different applications may be hosted on different servers. For example, in this figure, vehicle system A connects to the MSP center server, MSP edge server 1 and MSP edge server 2A. Vehicle system B connects with MSP edge server 2A, 2B and the MSP center server.

The objective of the mobility service instance selection service will vary, depending on each service scenario. Information such as vehicle geolocation, access network topology, server load, network performance and policy may be contributed as part of the selection process. The function that the mobility service instance selection service performs is to collect, process and distribute information about the available mobility service instances, enabling the applications within the vehicle system to connect to the most appropriate application server instance.

An AECC system may have a highly dynamic network topology. For this reason, the use of names (such as FQDN) is more flexible than endpoints using IP addressing. A naming scheme is therefore necessary as part of an overall resource scheme within the AECC system that the mobility service instance selection service can then utilize.

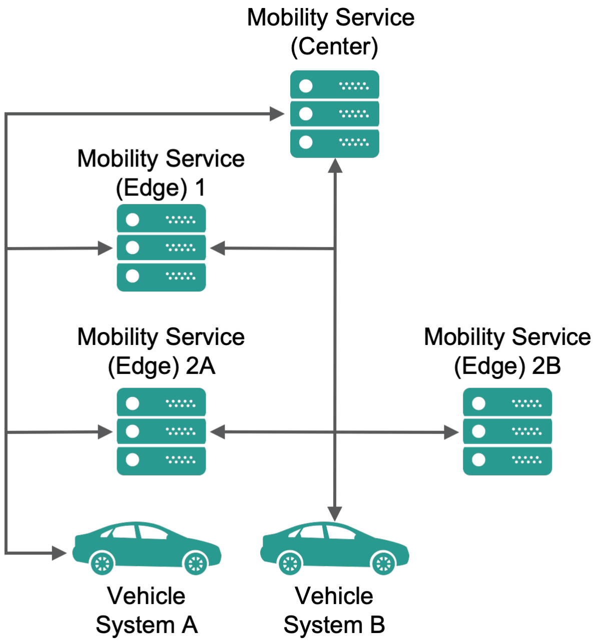
Mobility Service (Center)
Mobility Service (Edge) 1
Mobility Service (Edge) 2A
Mobility Service (Edge) 2B
Vehicle System A
Vehicle System B

Figure 23. Topology of hierarchical distributed computing, with lines and arrows showing the possible data exchange paths between all entities.

3.2.2 Potential Solutions

There are four solutions to the issue, as described in the following sections.

- Solution 1 –Cellular Network-based Mobility Service Instance Assignment
- Solution 2 –IP Network-based Mobility Service Instance Assignment
- Solution 3 –Mobility Service Instance Assignment by a Selection Function
- Solution 4 –Vehicle System-based Mobility Service Instance Assignment
- Solution 5 – Combination with Load Balancer for Mobility Service Instance Selection

The solutions are based on the entities that will select the target mobility service instance. The solution mapping is shown in Figure 24.

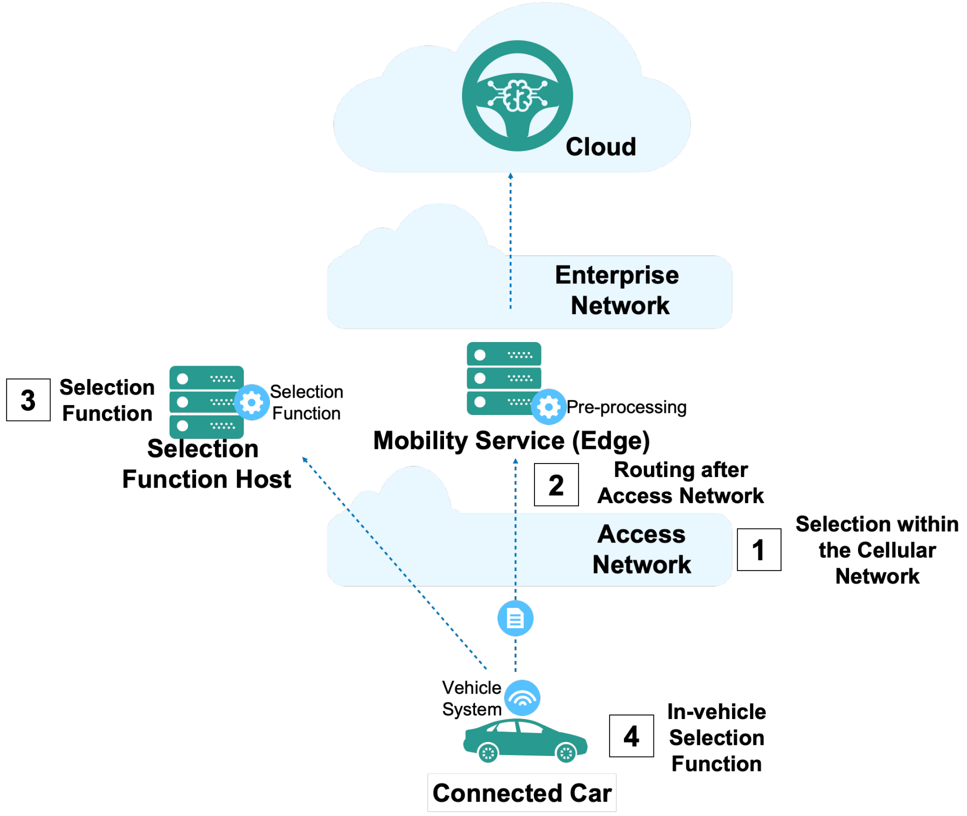

Figure 24. Mapping of potential solutions for mobility service instance selection.

3.2.2.1 Solution 1 – Cellular Network-based Mobility Service Instance Assignment

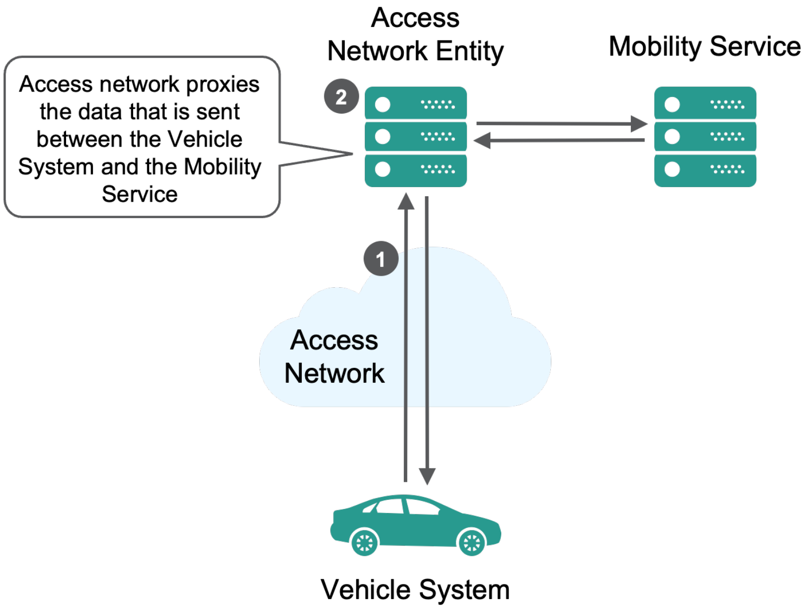
Access Network Entity
Mobility Service
Access network proxies the data that is sent between the Vehicle System and the Mobility Service
2
1
Access Network
Vehicle System

Figure 25. Cellular network-based mobility service instance assignment.

The location of a vehicle that connects to the cellular network may be known to a certain degree by the cellular network. Cellular networks also have SCEF (LTE) and NEF (5G) components that enable communication with the application instance. In this approach, the cellular network entity becomes a proxy and a control agent for communication between the vehicle system and the mobility service instance; for applications that are dependent on the vehicle system's movement, the vehicle system is oblivious to the mobility service instance assignment procedure. The approach assumes that the cellular network operator and the MSP have mutual agreement on how the assignment should occur.

1) The vehicle system sends data through the cellular network. This enables the cellular network to identify the vehicle's location through the base station to which the vehicle is connected.
2) The cellular network entity chooses the most suitable mobility service instance based on the agreement with the MSP and other criteria such as current server load. This process includes the selected hostname resolving.
3) The vehicle connects and is able to communicate with the target mobility service instance.
4) Response from the mobility service instance is then routed back to the vehicle system through the cellular network.

3.2.2.2 Solution 2 – IP Network-based Mobility Service Instance Assignment

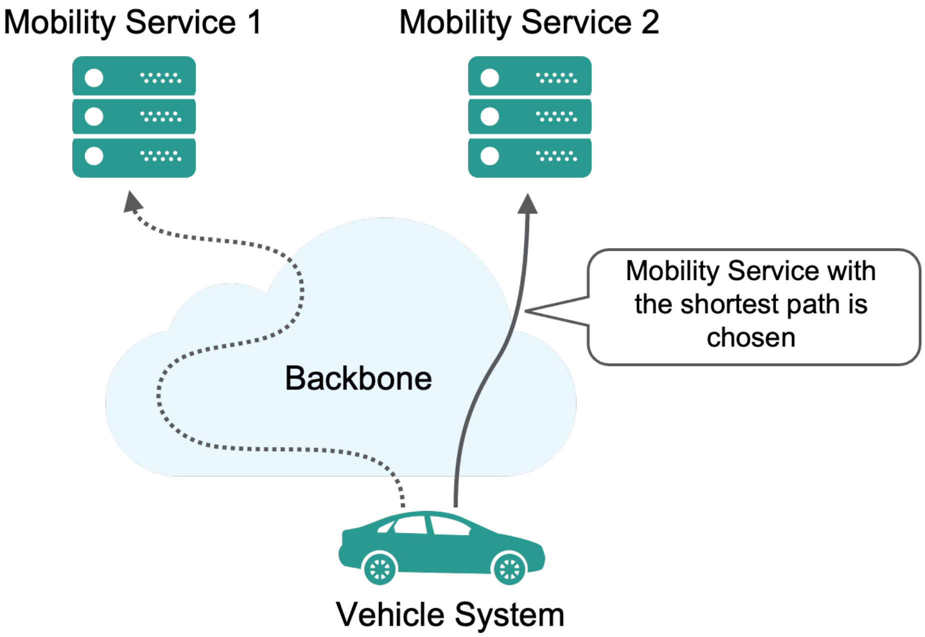

Figure 26. IP network-based mobility service instance assignment.

In this approach, both the mobility service instance and the vehicle system do not have any logic to decide upon the appropriate mobility service instance. The routing scheme may leverage IP anycast, so that traffic from the vehicle system will be forwarded by routers within the IP network to the mobility service instance with the shortest path. Application instances are deployed in a distributed manner, with application instances being provisioned with predetermined IP addresses. No selection function takes place, and instead network topology-based routing to the mobility service instance is used instead. In this approach, all required information is located in the application layer and not shared with the network, thus making it agnostic regarding the access network.

3.2.2.3 Solution 3 – Mobility Service Instance Assignment by a Selection Function

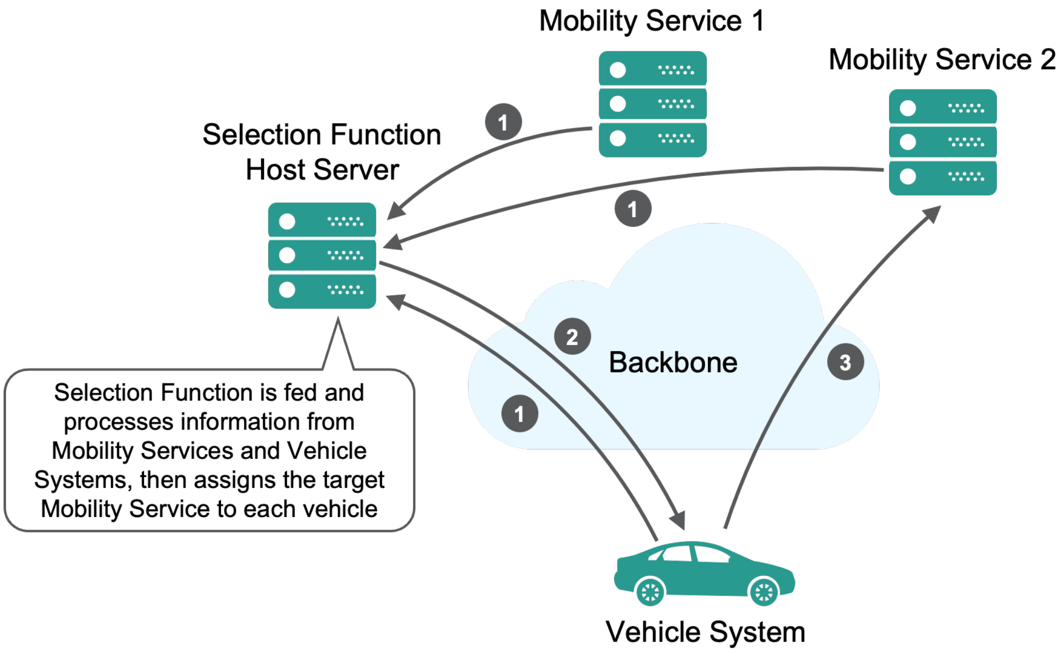
Mobility Service 1
Mobility Service 2
Selection Function Host Server
1
1
2
Backbone
3
1
Selection Function is fed and processes information from Mobility Services and Vehicle Systems, then assigns the target Mobility Service to each vehicle
Vehicle System

Figure 27. Selection function-based mobility service instance assignment.

In this approach, a selection function receives information from mobility service instances and the vehicle system. The selection function then processes the information and tells the vehicle system which mobility service instance to use, allowing the vehicle system to initiate a session with the selected mobility service instance. For example, the selection function could be implemented by a DNS server, where the algorithm runs while resolving a host name (Figure 27), although implementation is possible independent of the DNS system. Specific configuration of the selection function may allow processing of information shared by vehicle systems and mobility service instances, including but not limited to geolocation and/or server health check. This approach is agnostic with respect to the access network.

1) The selection function accepts information from mobility service instances and vehicle systems.
2) The selection function executes the selection algorithm and assigns the target mobility service instance.
3) The vehicle system connects to the target mobility service instance.

Figure 28 shows an example of implementing edge server assignment by a selection function using DNS.

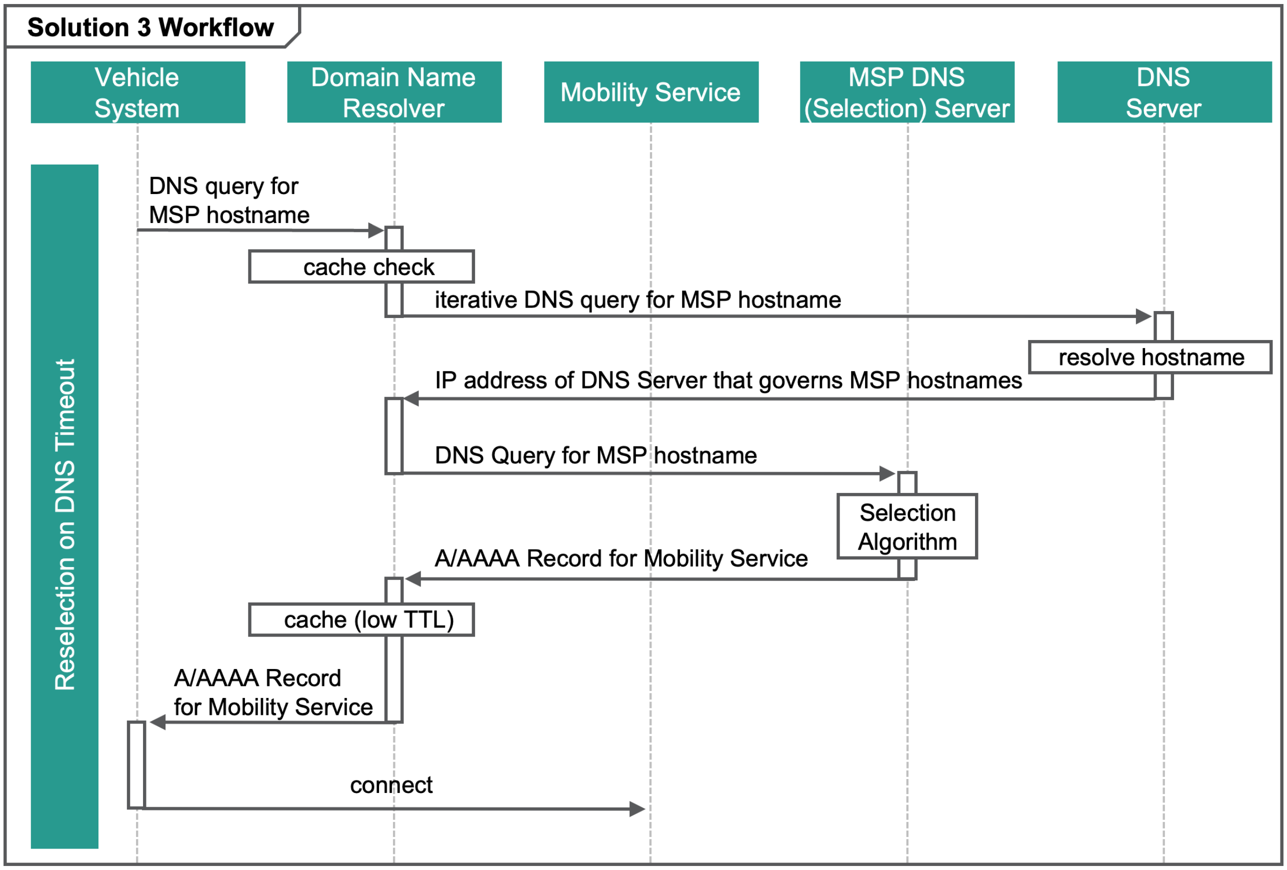
Solution 3 Workflow
Vehicle System
Domain Name Resolver
Mobility Service
MSP DNS (Selection) Server
DNS Server
Reselection on DNS Timeout
DNS query for MSP hostname
cache check
iterative DNS query for MSP hostname
resolve hostname
IP address of DNS Server that governs MSP hostnames
DNS Query for MSP hostname
Selection Algorithm
A/AAAA Record for Mobility Service
cache (low TTL)
A/AAAA Record for Mobility Service
connect

Figure 28. Example of an implementation of Solution 3.

Notes on the implementation:

□ Due to the vehicle system's movement within the environment, mobility service instance reselection may be required. To enable responsive mobility service instance reselection, appropriate DNS cache timeouts should be selected.

□ When an application is deployed in a manner such that it can cover only a certain geographic region, the selection function will need the vehicle's geolocation information. The region or area supported by a particular application instance may differ based on each service or as a result of the computing resources assigned to an application instance. This may need to be taken into consideration with respect to the DNS query.

□ The selection sequence takes into account that the resolved IP address may be an address of a load-balancer, thus triggering Solution 5. When using a 5GS, additional functions may be needed to enable the correct edge selection when coupled with the offloading function discussed in the previous key issue, Edge Data Offloading.

3.2.2.4 Solution 4 – Vehicle System-based Mobility Service Instance Assignment

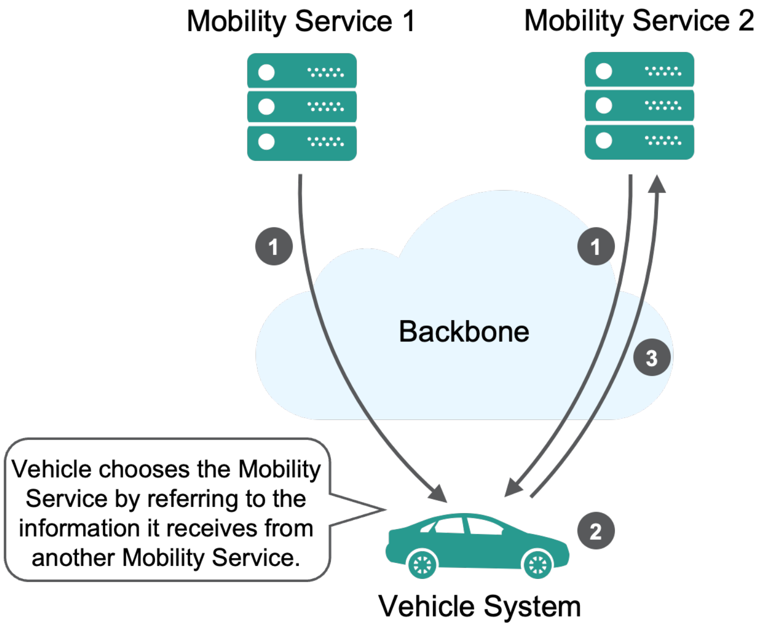
Mobility Service 1
Mobility Service 2
1
1
Backbone
3
Vehicle chooses the Mobility Service by referring to the information it receives from another Mobility Service.
2
Vehicle System

Figure 29. Vehicle system-based mobility service instance selection.

This solution can be combined with the selection function solution, allowing mobility service instance assignment without the vehicle system having to share sensitive information, such as vehicle geolocation. In this approach, the vehicle system will choose its mobility service instance. The vehicle system may select the appropriate mobility service instance based on in-vehicle information such as physical vehicle location and/or additional information provided by potential mobility service instances. This approach is agnostic with respect to the access network.

1) The vehicle system requests information from mobility service instances.
2) Based on the information received by the vehicle system, the vehicle system selects a mobility service instance.
3) The vehicle system connects to the target mobility service instance.

Figure 30 shows an example of implementing the combination of a selection function and vehicle-based assignment.

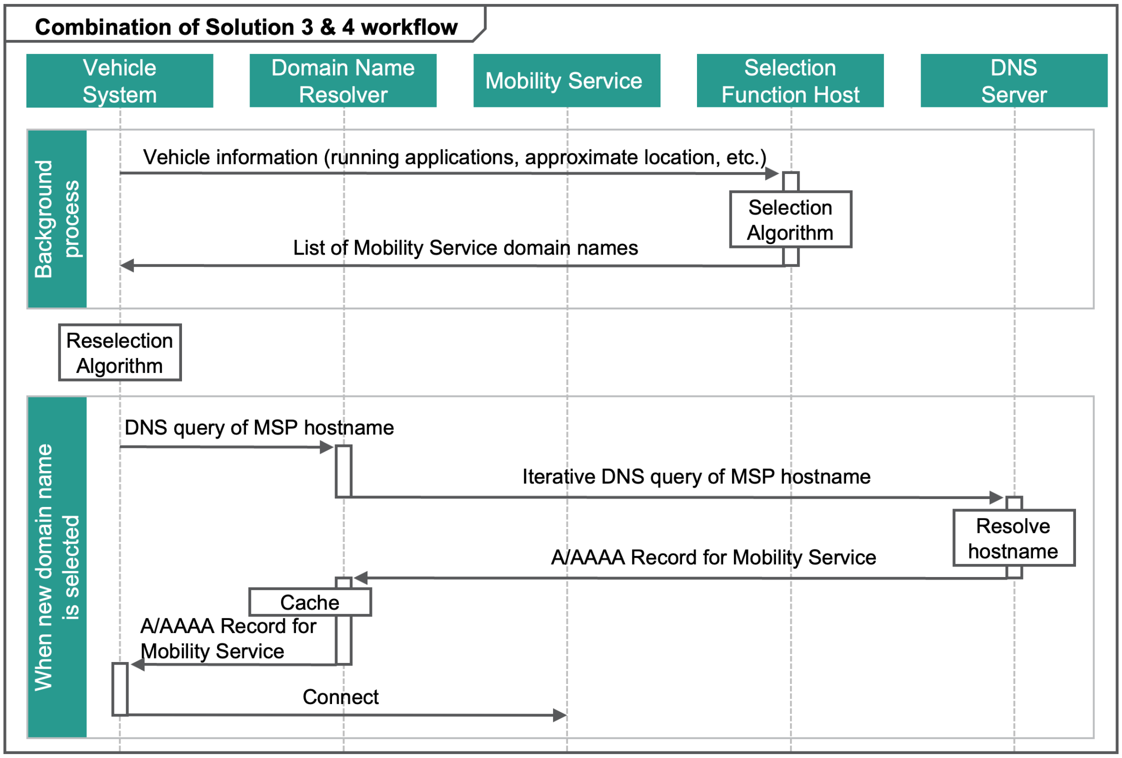
Combination of Solution 3 & 4 workflow
Vehicle System
Domain Name Resolver
Mobility Service
Selection Function Host
DNS Server
Background process
Vehicle information (running applications, approximate location, etc.)
Selection Algorithm
List of Mobility Service domain names
Reselection Algorithm
When new domain name is selected
DNS query of MSP hostname
Iterative DNS query of MSP hostname
Resolve hostname
A/AAAA Record for Mobility Service
Cache
A/AAAA Record for Mobility Service
Connect

Figure 30. Example of an implementation of the combination of Solution 3 and Solution 4.

3.2.2.5 Solution 5 – Combination with Load Balancer for Mobility Service Instance Selection

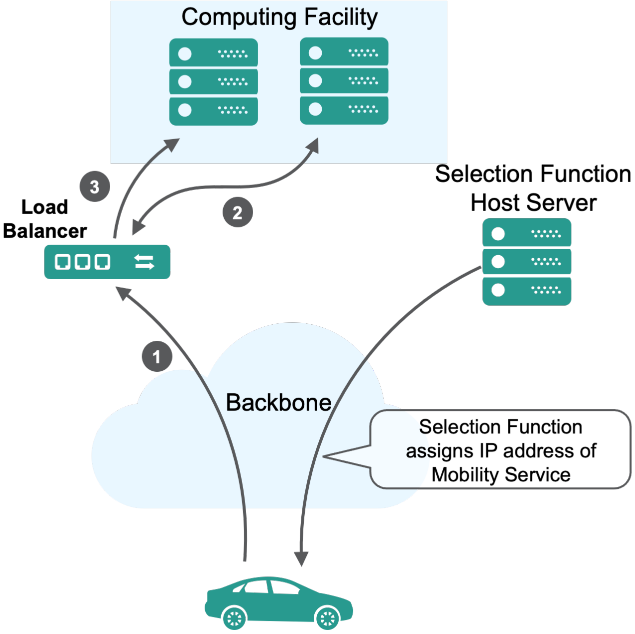
Computing Facility
3
Load Balancer
2
Selection Function Host Server
1
Backbone
Selection Function assigns IP address of Mobility Service

Figure 31. Example illustrating a load balancer coupled with selection on a server.

This method combines one of the selection methods with an application-aware load balancer in order to determine which application instance should be used by the vehicle system. One must be aware that the MSP edge server is not a single server but rather a set of servers in one or more data centers composed of the necessary functionality to support the various service scenarios. In the diagram above, the vehicle system will request IP address resolution for a particular service, with the result being an appropriate IP address that happens to be allocated to a load-balancer device.

1) The vehicle system initiates a connection toward the load balancer.
2) The load balancer determines which mobility service instance to use.
3) The load balancer will manage the session on behalf of the vehicle system toward the selected application instance within the MSP data center.

3.2.2.6 Parameters for Mobility Service Instance Assignment

The following parameters may be used for mobility service instance assignment.

- IP Ping **Round Trip Time (RTT) –** The RTT between the vehicle system and reachable mobility service instances.
- **Completion Rate –** The quality of this parameter is inversely proportional to the number of packet losses and timeouts between the vehicle system and the mobility service instance.
- **Hops –** Number of hops needed to route the data between the vehicle system and the mobility service instance.
- **Vehicle System Physical Location –** To address localized contents/process, the vehicle system's physical location may be required. Attention to security and privacy considerations is required.
- **Request SLA –** By identifying the dataflow, the system is able to identify the required SLA of the request, including the required time needed to complete the processing of a particular data flow.
- **Server Turnaround Time –** The time needed for an application in the mobility service instance to complete a process queried by the vehicle system.
- **Server Load –** The load of a mobility service instance; this might include CPU and memory utilization.

3.2.2.7 Considerations across Key Issues

In current deployments of LTE and WLAN networks, when dynamic IP address assignment and configuration are used, the network typically provisions the IP address of a DNS server. The DNS server may be located close to the network edge and configured in a manner so as to provide results that are aware of resources that are close to it. Conversely, a DNS server may reside in a remote network and therefore be unaware of resources located within the local network. In 5G, an MNO's DNS Server may serve clients from multiple anchor points.

Consider now the combination of a DNS-based solution with an uplink classifier-based solution, which is the preferred edge data offloading solution for a 5GS. As the user has multiple possible anchor points of presence that can be used based on IP header filtering (typically destination IP address), the cellular network must ensure that a DNS query is sent to a domain name resolver at a point of presence that is deemed feasible for the corresponding FQDN and UE location in the network.

As a solution, the mobile network operator operating the respective cellular network should provide a central DNS stub resolver that, when receiving a DNS query, forwards this query to an appropriate domain name resolver. This action is based on the tracking area (looked up using standard 5G core functionality) of the UE, and potentially on additional knowledge related to the FQDN, such as rules about feasibility of different breakouts for specific domain names. In addition, the DNS stub resolver may aggregate multiple records in the reply.

3.2.3 Conclusions

Mobility service instance selection schemes are heavily influenced by the requirement of the mobility services and how they are deployed within the architecture. Recommendations on the technologies to be used for mobility service instance selection are as follows:

- Solution 3 is preferable for deployments in the near future, since this solution does not involve major modifications to the vehicle system. Furthermore, it minimizes the customization effort for the access network. Dedicated DNS servers (MSP DNS servers) that are authoritative for the corresponding DNS zone(s) are recommended to be deployed for mobility service instance selection, to enable seamless integration with existing deployed systems, due to their wide adoption on the internet, in access networks and on existing clients.
- Solution 4 allows the vehicle system to select mobility services without the vehicle system having to share information related to the vehicle. This solution also allows the mobility service instance selection function to give possible options to the vehicle and the final decision to be made by the vehicle. This option is also transparent to the access network. A specific selection module must be implemented within each vehicle system.
- Solution 1 allows a deployment that is transparent to the vehicle system and mobility services.
- In conclusion, Solution 3 is recommended as a baseline solution, preferably reusing existing DNS functionality and infrastructure, while both Solutions 1 and 4 are feasible to offer enhanced functionality for the mobility service instance selection process, which may be considered when such requirements are called for by a mobility service.

3.3 Vehicle System Reachability

3.3.1 Key Issue

In the AECC distributed computing architecture, mobility service instances – center or edge – are required in many use cases to send data to the vehicle system. However, it is challenging for the mobility service instances to effectively reach the vehicle system.

1) The IP anchor point changes.
2) The IP mapping changes due to network functions such as an NAT/NAPT timer’s expiration, resulting in a change of the global IP address.
3) A service outage occurs when the vehicle system moves into an area without network coverage.
4) Handover between different access networks occurs.

All these issues can cause a vehicle system IP address change that results in mobility service instances not being able to reach the vehicle system. Therefore, the AECC system needs to deploy a specific mechanism to ensure the reachability of a vehicle system according to service requirements.

Note 1: depending on the service requirements, the vehicle system reachability issue could be handled differently by the application layer or by both the application and network layer.

Note 2: both IP and non-IP based solutions shall be considered for this key issue.

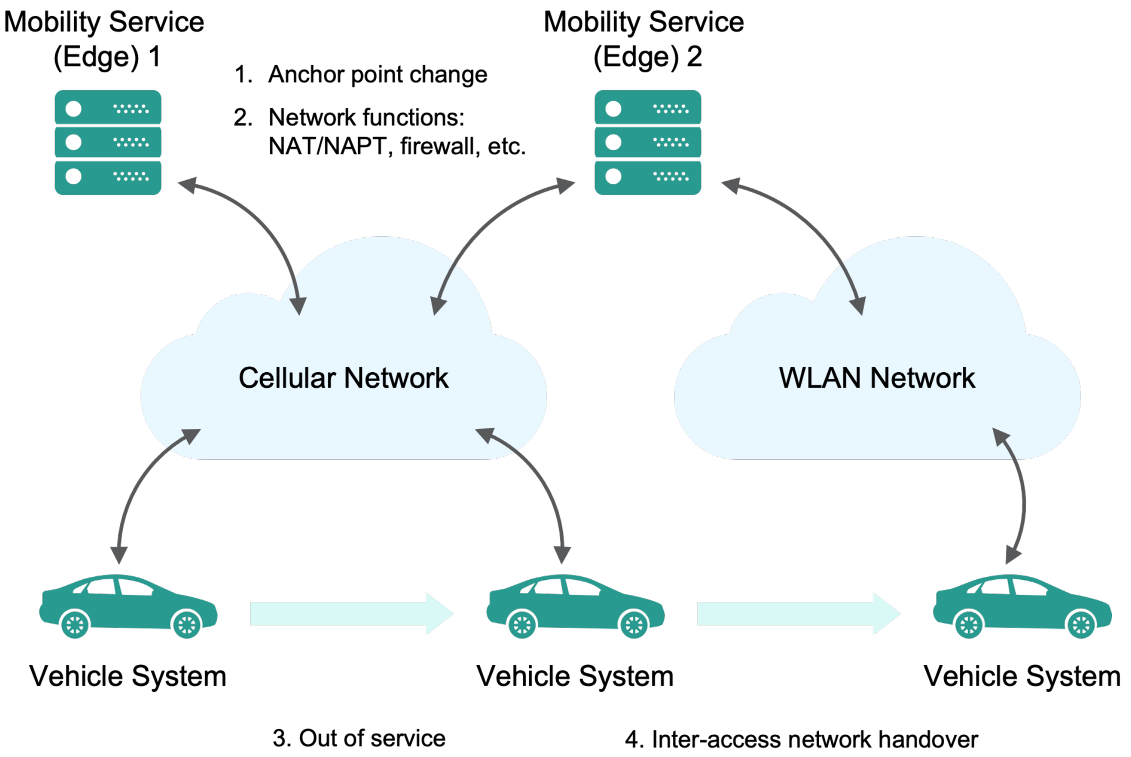
Mobility Service (Edge) 1
1. Anchor point change
2. Network functions: NAT/NAPT, firewall, etc.
Mobility Service (Edge) 2
Cellular Network
WLAN Network
Vehicle System
Vehicle System
Vehicle System
3. Out of service
4. Inter-access network handover

Figure 32. Typical causes of the vehicle system reachability issue.

3.3.2 Potential Solutions

Three potential solutions are investigated in the following sections.

□ SMS Push
□ Push Notifications
□ Vehicle System Triggering via Network Exposure Function

3.3.2.1 Solution 1 – SMS Push

SMS Push is an SMS “call-in” trigger message common in many current Over-the-Air (OTA) software delivery systems. The trigger message is sent to the target vehicle system by a campaign manager application. The application maintains links to a database operated by the vehicle manufacturer that contains the IMEI number of each vehicle system within the fleet. Database fields enable the vehicle manufacturer to identify subsections of the vehicle fleet based on parameters such as vehicle type, country and so on. When the SMS trigger message is received, the vehicle system initiates the appropriate application. The application will then establish a connection to the OTA delivery system. Once a vehicle system connects, it is marked as having received the message and taken the action. This allows the OTA system to identify those vehicles that have not yet connected.

Since the application in the vehicle system initiates the connection into the OTA delivery system, the IPv4/IPv6 address of the vehicle system will be obtained when the connection is established.

3.3.2.2 Solution 2 – Push Notifications

Push notifications are a common method for maintaining connectivity used in mobile consumer device platforms, such as smartphone applications.

A push notification is a message that is “pushed” automatically from a backend server or application to remote clients. These notifications are sent from the application to a remote server, which acts as an intermediary. Each client application needs to be registered with the remote server using a unique key or UUID. The remote server then sends the message against the unique key and delivers the message to the client application via an agreed client/server protocol such as HTTP or XMPP.

The vehicle system’s IPv4/IPv6 address is not required since the backend server or application communicates with the remote server, and the client-side application in the vehicle system registers with the remote server.

3.3.2.3 Solution 3 – Vehicle System Triggering via Network Exposure Function

Network exposure functions, such as SCEF (defined in 3GPP TS 23.682) and NEF (defined in 3GPP TS 29.522), specify various network APIs for third parties. The vehicle system could be triggered via a control plane message to establish its IP connection to the mobility service instances with its currently used IP address.

One solution example demonstrated in Figure 33 shows how the oneM2M framework leverages the exposure functions of cellular networks [8]. In this example, the oneM2M function in the mobility service instance will continuously maintain the ID-IP binding between the vehicle oneM2M ID and its IP address. The binding can be updated periodically or by an event that uses the cellular network exposure function to trigger the oneM2M function in the vehicle system for updating its IP address. The applications using the oneM2M framework in the mobility service instance will only need to know the unique ID to wake and reach the vehicle system.

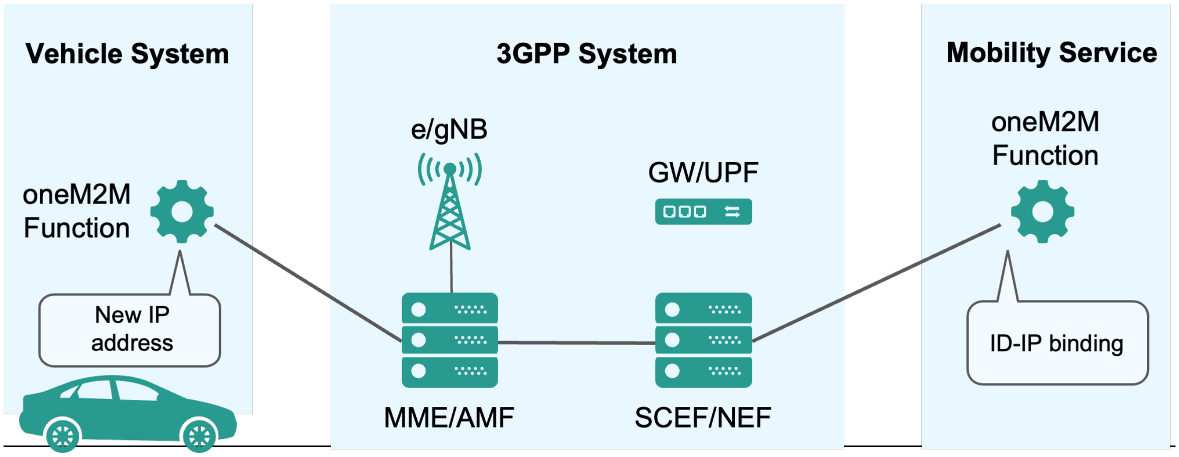

Figure 33. Examples of network exposure functions.

3.3.3 Conclusions

Recommendations on the technologies to be used for vehicle system reachability are as follows.

- Solution 2 is preferable for supporting both cellular network and WLAN access types, which is an architectural requirement in the AECC system. It is recommended to deploy this solution as an AECC system function, but it is also flexible enough to be implemented by a third-party application that is outside the AECC system.

- Solution 1 could be a fallback solution when the cellular network is the only access to be used for push services, since Solution 1 already has good adoption in many existing industry sectors, such as automotive and other IoT systems.
- Solution 3 would be promising to replace Solution 1 in cellular networks that support the SCEF/NEF, due to the advantages of less vehicle system complexity and better extensibility as compared to Solution 1.

3.4 Access Network Selection

3.4.1 Key Issue

As shown in Figure 34, the vehicle system is expected to use a mix of different wireless access technologies, including cellular and WLAN, to connect to mobility service instances. It may be preferable for the vehicle system to use multiple access networks simultaneously to access increased bandwidth or to improve reliability. For example, a vehicle system may be traveling through an area where service coverage from network operators improves and degrades. How should the vehicle system adapt to the changing communications environment? At the same time data flows with different QoS requirements or from different applications may be required to go through different access networks in order to meet AECC system service requirements. Besides the access network status, information such as policies, service requirements, network connectivity and the vehicle system’s movement can be considered as part of the process. Therefore, a mechanism is needed to enable the vehicle system or the AECC system to select from multiple access networks and steer traffic over different connections. The elements involved in this process are access networks, the vehicle system and mobility service instances.

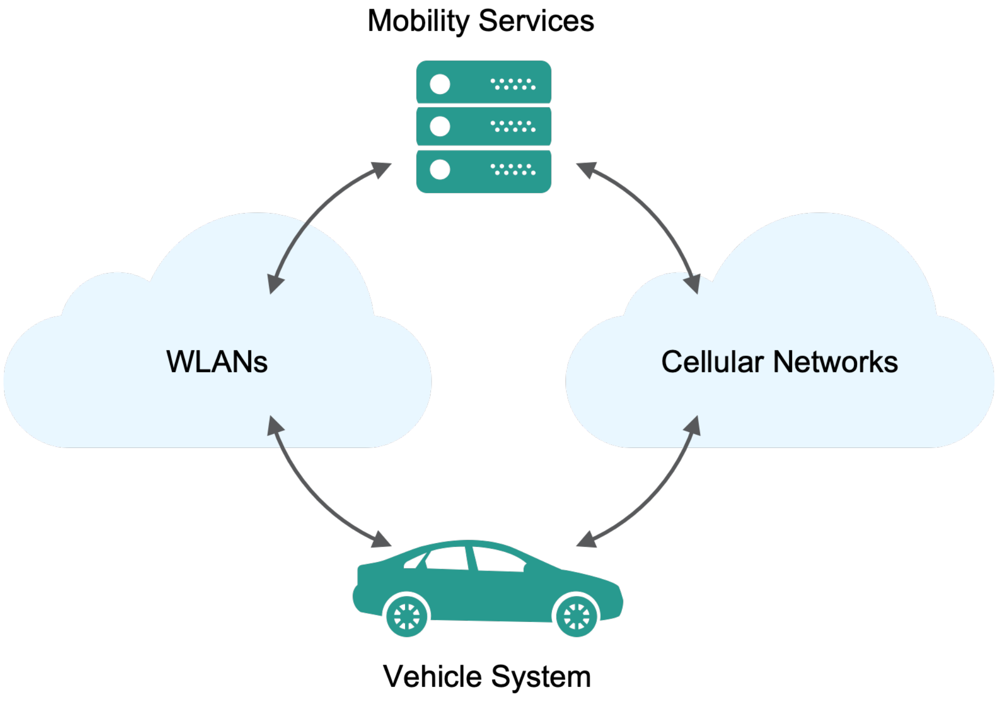

Figure 34. Access network selection.

3.4.2 Potential Solutions

Access network selection can be divided into two steps: connection selection and traffic steering. Connection selection focuses on how to select one or multiple connections based on the end-to-end system capabilities and other information from the vehicle system, the cellular network(s), the WLAN(s) and mobility service instances. Examples of relevant information are wireless signal strength, application requirements and the movement of the vehicle system through the environment. Traffic steering focuses on how to steer traffic to multiple connections and enable different applications to use different connections simultaneously. Access network selection can be triggered periodically or by events.

Solutions for access network selection are listed below.

Solutions for connection selection:

- Solution 1 – Vehicle System-based Solutions for Connection Selection
  - Solution 1.1 – Application layer solution on the vehicle system
  - Solution 1.2 – System layer solution on the vehicle system
- Solution 2 – Mobility Service Instance-based Solutions
  - Solution 2.1 – Application layer solutions on the mobility service instance
  - Solution 2.2 – System layer solutions on the mobility service instance
- Solution 3 – Access Layer Solutions
  - Solution 3.1 – ANDSF
  - Solution 3.2 – P-GW level convergence based on S2a/S2b interfaces

Solutions for traffic steering:

- Solution 4 – System Layer Solutions
  - Solution 4.1 – MPTCP/MPQUIC
  - Solution 4.2 – Generic Multi-Access (GMA)/Multi-Access Management Services (MAMS)
- Solution 5 – Access Layer Solutions
  - Solution 5.1 – LTE WLAN Aggregation (LWA)
  - Solution 5.2 – LTE WLAN radio-level integration over IPSec tunnel (LWIP)
  - Solution 5.3 – Access Traffic Steering, Switch and Splitting (ATSSS)
  - Solution 5.4 – Multi Radio Dual Connectivity (MR-DC)

The first section below provides the solution overview. Afterward, all solutions are defined one by one.

3.4.2.1 Solution Overview

As shown in Figure 35, the basic approach for both connection selection and traffic steering takes multiple inputs from the vehicle system, mobility service instances and access networks to make decisions about selecting one or multiple connections, or steering traffic over the selected connections.

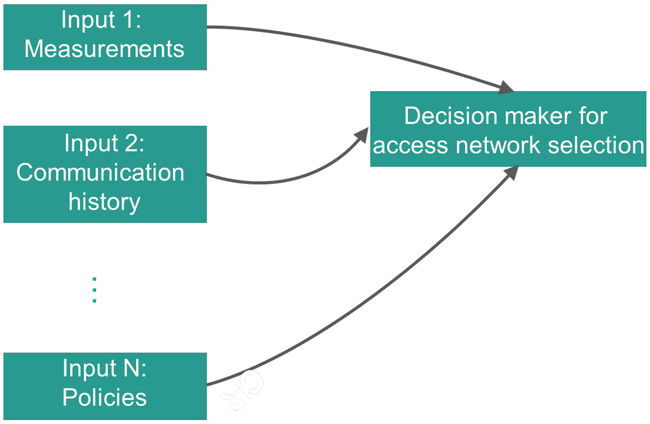

Figure 35. Basic approach for access network selection.

The inputs for access network selection may include:

- Policy: this refers to the policies used to govern access network selection, such as
  - A condition regarding channel state for selecting an access network. For example, a threshold of the SINR can be defined to decide when to switch between access networks.
  - A condition regarding the vehicle system's location for selecting an access network. For example, the vehicle system may switch to a WLAN once it arrives home.
  - A condition regarding service or traffic types for selecting an access network.
  - Operational policies such as using a primary access network if it is available.
- Data: this refers to the data or information used to assist access network selection (note that the inputs are not mandatory and depend on availability), such as
  - Radio link measurements about different RATs, such as signal strength, SINR
  - Communication history
  - Access network charging information
  - Network status, such as congestion or Round Trip Time (RTT)
  - Location information for the vehicle system or mobility service instances
  - Service-related information, such as service type, data rate, traffic behavior, traffic statistics, priority
  - Application context, such as the user's preference

As shown in Figure 36, the access network selection assistance function can be located within different layers on the vehicle system or the mobility service instances:

- The access layer generally consists of cellular and WLAN networks, including both radio access networks and core networks, and other infrastructure that provides connectivity between the vehicle system and mobility service instances.
- The system layer generally includes a function layer, a platform layer, an infrastructure layer and computing, network and storage on the vehicle system and mobility service instances.
- The application layer includes the applications running on the vehicle system and mobility service instances.

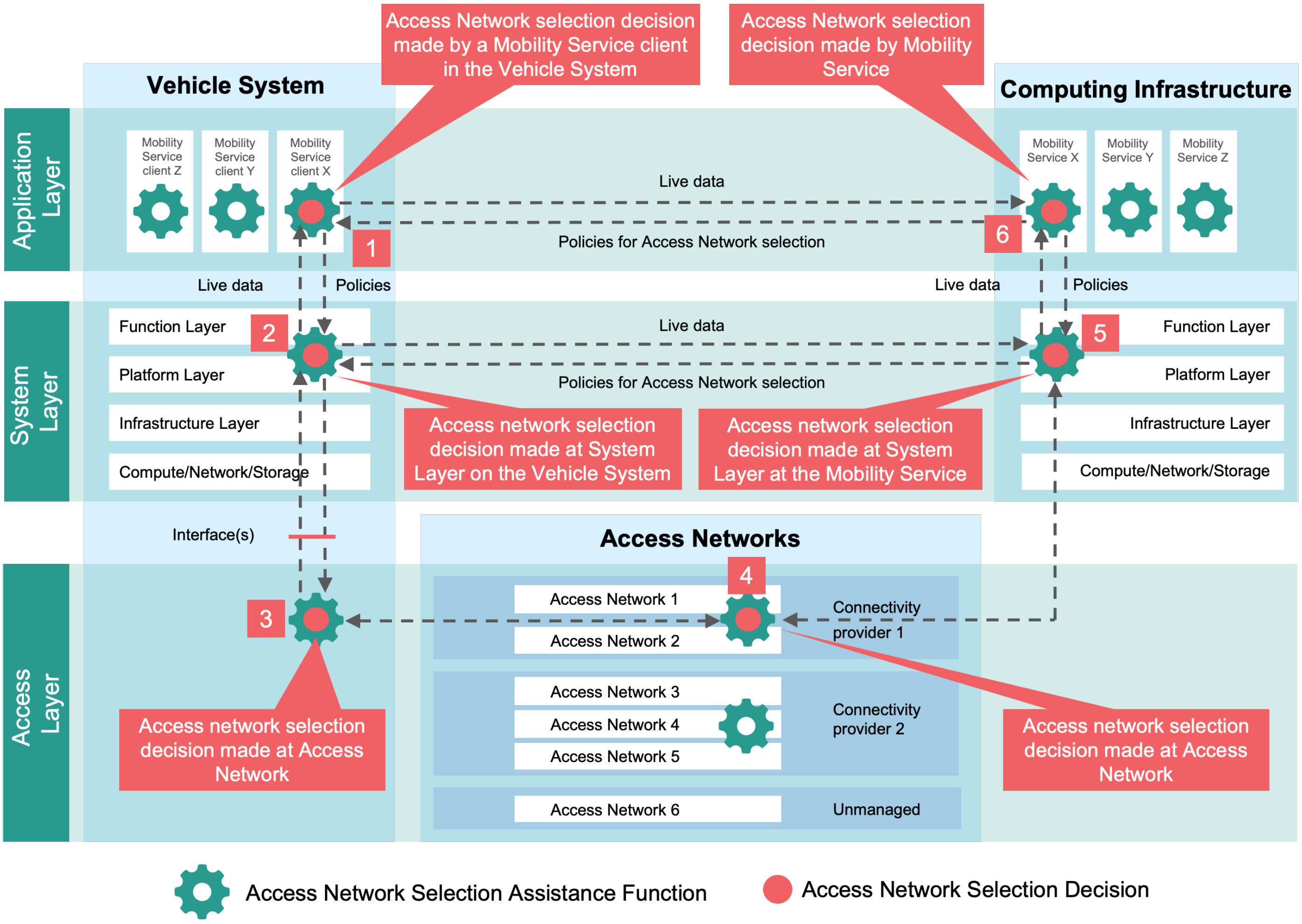
Access Network selection decision made by a Mobility Service client in the Vehicle System
Access Network selection decision made by Mobility Service
Vehicle System
Computing Infrastructure
Application Layer
Mobility Service client Z
Mobility Service client Y
Mobility Service client X
Mobility Service X
Mobility Service Y
Mobility Service Z
Live data
Policies for Access Network selection
1
6
Live data
Policies
Live data
Policies
System Layer
Function Layer
Platform Layer
Infrastructure Layer
Compute/Network/Storage
2
5
Live data
Policies for Access Network selection
Access network selection decision made at System Layer on the Vehicle System
Access network selection decision made at System Layer at the Mobility Service
Interface(s)
Access Networks
4
3
Access Layer
Access Network 1
Access Network 2
Connectivity provider 1
Access Network 3
Access Network 4
Access Network 5
Connectivity provider 2
Access Network 6
Unmanaged
Access network selection decision made at Access Network
Access network selection decision made at Access Network
Access Network Selection Assistance Function
Access Network Selection Decision

Figure 36. Possible information sharing and decision making for access network selection.

Figure 36 illustrates the possible information exchange between functions that may assist the selection process. The application layer may provision policies to the system layer, which again may provision policies to the access layer. Access network selection assistance functions outside the vehicle system may provision policies to the vehicle system, within the respective layer. In addition, the access layer may share live data with the system layer, and the system layer may share live data with the application layer. The vehicle system may also share live data with access network selection assistance functions outside the vehicle system, within the respective layer. For example, this could include metrics relating to throughput. For a practical solution, only a subset of the mentioned information is shared, depending on where the selection algorithm is instantiated and the requirements for access network selection.

While access network selection in the system and application layers deals with the selection of access networks exposed by the modem in the vehicle system (this could be a single access network per connectivity provider), selection in the access layer deals with the selection of access networks by a single connectivity provider, and the two selection processes may be combined.

Figure 36 also illustrates how the described layers relate to the reference architecture. The key point is that on both the vehicle system and mobility service instance sides, access network selection assistance functions in different layers can work together to enable different granularities of access network selection. Placing the application layer selection assistance function will be application-specific, while placing the selection assistance function in the system layer is vehicle system-specific.

The decision on access network selection includes the decision on connection selection and the decision on traffic steering. The decision on connection selection may be made at a different layer from where the decision on traffic steering is made. However, the design of the whole system shall make sure that the mechanisms of connection selection and traffic steering can work together.

3.4.2.2 Solutions for Connection Selection

Connection selection for links includes two stages:

- Stage 1: initial connection selection for selecting a connection without any existing connection
- Stage 2: connection reselection for selecting a connection with one or more existing connections

Initial connection selection would be largely based on in-vehicle information and provisioned policies. Connection reselection could be conducted by a mobility service instance or the vehicle system, or assisted by a cellular network with the assumption that one or more connections is successfully established.

Types of connection selection include:

- Type 1: only one connection is selected at one time.
- Type 2: multiple connections are selected and can be activated simultaneously.

Type 1 connection selection generally can be based on policies such as using a WLAN if available and using a cellular network otherwise. It can also be provided by Access Network Discovery and Selection Function (ANDSF) policies, which will be introduced in Solution 3.1 in this section.

Type 2 connection selection could consider the end-to-end system capabilities of traffic steering. For example, the vehicle could select the following combinations of access networks: independent access networks such as primary cellular, secondary cellular and WLAN by different operators, or access networks that can support some interworking features, such as a cellular network and a WLAN connected to the same core network. ANDSF also defines policies for selecting multiple access networks.

The connection selection can be made at the application, system or access layers on the vehicle system or mobility service instances. The decision on selecting a connection at different layers will result in selecting one or multiple physical links or logical links. However, a logical link can be mapped to a physical link, and eventually to an access network.

In Table 1, the solutions for connection selection are summarized based on where the connection selection decision is made, conditions and requirements, and the selection types and stages as defined above. The details of each solution are introduced afterward.

Table 1. Summary of solutions for connection selection.

Solutions	Layers where decision is made	Option number in Figure 36	Assumptions and requirements	Selection type and stage
Solution 1.1 -- Application layer solution on the vehicle system	At the application layer on the vehicle system	1	The northbound interface from the system layer to applications and the southbound interface to the access network may be required on the vehicle system.	Stage 1 and Stage 2 for Type 1 or Type 2
Solution 1.2 – System layer solution on the vehicle system	At the system layer on the vehicle system	2	The northbound interface from the system layer to applications and the southbound interface to the access network may be required on the vehicle system.	Stage 1 and Stage 2 for Type 1 or Type 2
Solution 2.1 -- Application layer solutions on the mobility service instance	At the application layer on the mobility service instance	6	The northbound interface from the system layer to applications and the southbound interface to the access network may be required on the mobility service instance.	Stage 2 for Type 1 or Type 2
Solution 2.2 -- System layer solutions on the mobility service instance	At the system layer on the mobility service instance	5	The northbound interface from the system layer to applications and the southbound interface to the access network may be required on the mobility service instance.	Stage 2 for Type 1 or Type 2
Solution 3.1 -- ANDSF	At the access layer	3	Both the vehicle system and the access network have to support ANDSF.	Stage 2 for Type 1 or Type 2
Solution 3.2 -- P-GW-level convergence based on S2a/S2b interfaces	At the access layer	4	Applied to the LTE network	Reselection for a WLAN for Type 2

Solution 1 – Vehicle System-based Solutions for Connection Selection

The vehicle system-based solutions apply to the scenarios where the vehicle system makes the decision on connection selection.

Solution 1.1 – Application layer solution on the vehicle system

Connection selection can be implemented as part of the application on the vehicle system that receives inputs from different layers, such as the application, system and access layers on both the vehicle system and mobility service instances, to make decisions, shown as Option 1 in Figure 36. For example, access networks could provide measurements of signal strength and an available network list or policies from an ANDSF to the connection selection function in the application in order to make a decision. In addition, the connection selection functions in the application layer on the mobility service instance can provide policies for connection selection. This solution can also integrate the user's preference to make a selection decision.

To facilitate this solution, it may be possible for the vehicle system to provide northbound interface APIs with connection selection functions at the application layer. Functions of the northbound interface may include the vehicle system providing information about the AECC system's capabilities in supporting accessnNetworks, and providing information (e.g., access network conditions, location) for the application to make a selection decision. The vehicle system could also provide a southbound interface to the access network. Functions of the southbound APIs may include access network measurement reports and the access network charging policy. However, the vehicle system may have no or limited support for the aforementioned capabilities.

Solution 1.2 – System layer solution on the vehicle system

The decision on connection selection can be made at the system layer on the vehicle system, which can get inputs from the connection selection assistance functions in the application and access layers on both the vehicle system and mobility service instances, shown as Option 2 in Figure 36. In this solution, the connection selection decision can be agnostic about applications.

To facilitate this solution, the vehicle system's system layer could provide a northbound interface to applications. Functions of the northbound APIs may include getting the policies or user preference for connection selection. The vehicle system could provide a southbound interface to the access network. Functions of southbound APIs may include access network measurement reports and the access network charging policies.

Solution 2 – Mobility Service Instance-based Solutions

Mobility service instance-based solutions apply to the scenarios where mobility service instances make the decision on connection selection once an initial connection has been established by the vehicle system. The selected access networks can be either cellular networks or WLANs. The mobility service instance-based solutions apply only to connection reselection and require capabilities from the mobility service instance to have related interfaces and get information about access networks.

Solution 2.1 – Application layer solutions on the mobility service instance

The decision on connection selection can be made at the application layer on the mobility service instance, which can get inputs from the connection selection assistance functions in the access layer and system layer on both the vehicle system and mobility service instances, shown as Option 6 in Figure 36. The information from the access network can be sent to the application layer or system layer on the mobility service instance via an application layer interface between the vehicle system and mobility service instance or via the southbound interface between the mobility service instance and the access networks.

A primary connection needs to be established for the communication between the vehicle system and mobility service instances**.** To facilitate this solution, the AECC system could provide northbound interface APIs to the application. Functions of the northbound APIs include providing information on the AECC system's capabilities in supporting access networks, providing information (e.g., access network conditions, location) for the application to make a selection decision. The AECC system shall provide a southbound interface to the access network.

Functions of southbound APIs may include providing access network measurement reports and the access network charging policy.

Solution 2.2 – System layer solutions on the mobility service instance

The decision on connection selection can be made at the system layer on the mobility service instance, which can make use of the connection selection assistance functions in the application and access layers on both the vehicle system and mobility service instances, shown as Option 5 in Figure 36. In this case, connection selection assistance functions at the system layer can be provided as standardized or proprietary APIs. A primary connection needs to be established for the communication between the vehicle system and mobility service instances for the vehicle system to provide information to the mobility service instances in order to execute the connection selection assistance function and notify the vehicle system of the result.

To facilitate this solution, the AECC system could provide northbound interface APIs to applications. Functions of the northbound APIs may include getting the policies or user preference on network selection. The connection selection assistance function may only be able to use “self-monitored” information such as traffic statistics. If the mobility service instances have interfaces to the access networks, information exchanged with the access network may be leveraged. The AECC system may provide a southbound interface to the access network. Functions of the southbound APIs may include providing access network measurement reports and the access network charging policy.

Solution 3 – Access Layer Solutions

In these solutions, connection selection is assisted by the cellular network, and the cellular network needs to support certain capabilities to enable access network selection.

Solution 3.1 – ANDSF

In EPS or 5GS, an ANDSF can provide an available network list or connection selection-related policies as specified in a 3GPP specification [10]. The decision on connection selection is made in the vehicle system based on the policies received from the ANDSF, shown as Option 3 in Figure 36. The input to the ANDSF is the UE’s profile, which specifies what output is expected from the ANDSF.

The UE may retain and use the access network discovery information provided by the ANDSF until new/updated information is retrieved. The ANDSF communicates with the UE over the S14 reference point, which is essentially a synchronization of an OMA-DM management object (MO) specific to the ANDSF.

This solution requires that the vehicle system can receive the ANDSF policy from the cellular MNO and that both the vehicle system and cellular networks support the ANDSF. This solution applies to both initial connection selection and connection reselection. The ANDSF does not provide RAN-level information but mostly provides information based on the vehicle system’s location.

Solution 3.2 – P-GW level convergence based on S2a/S2b interfaces

In EPS, S2a/S2b can provide Core Network (CN)-based interworking between cellular and WLAN networks. The S2a or S2b is the interface between the P-GW and a trusted Non-3GPP IP access or an untrusted Non-3GPP IP access to offload traffic to a non-3GPP access network such as a WLAN. The decision on connection selection is made by the access network based on policies such as user preference, RAN-level policies or WLAN service provider policies, shown as Option 4 in Figure 36.

Connection selection between 3GPP access and a WLAN is supported using an ANDSF or using RAN rule procedures without an ANDSF. As in Solution 3.1, an ANDSF can provide a list of available WLAN networks and related information to facilitate discovery and connection establishment. However, it does not contain information at RAN level, such as signal strength. When the vehicle system has valid 3GPP subscription credentials (i.e., a valid USIM) and WLANSP policies, the vehicle system can perform WLAN selection based on WLANSP policies, the applicable user preferences and the corresponding procedures as specified in a 3GPP specification [9]. This solution requires that the access networks integrate with the cellular core network, and it only applies to connection reselection for a WLAN network.

3.4.2.3 Solutions for Traffic Steering

Based on how traffic steering is supported, the solutions can be classified as:

- Non-seamless multi-access mode: in this case, multiple-access networks operate independently and no coordination is provided.
- 3GPP seamless multi-access: in this case, traffic steering relies on the 3GPP network to provide seamless multi-access service. Convergence between the cellular network and WLAN can be provided at CN level or RAN level. Generally, two access networks are supported; i.e., one cellular network and one WLAN. A single IP address may be used, independent of the underlying access networks.
- Non-3GPP seamless multi-access: in this case, a higher-layer solution needs to be supported for cellular and WLAN network convergence. Convergence can be provided to more than two access networks simultaneously. A single IP address may be used, independent of the underlying access networks.

Traffic steering enablers are listed below:

- **Enabler 1 – APN**
  One or more APNs can be allocated to an application to identify different flows as agreed with the operators. In this case, an application may be presented with endpoints (such as next-hop IP addresses) that can be used to place traffic onto a particular APN. This technology requires agreement with the MNO to assign APNs and applies only to cellular networks for non-seamless multi-access mode.
- **Enabler 2 – Socket Binding**
  Socket binding can bind a socket to a specific network interface IP address. For example, if an application opens a socket for the TCP protocol and explicitly binds it to the WLAN interface, then the socket sends and accepts data only from the WLAN interface. This allows an application to transfer data only when the WLAN is available. This solution applies to both the cellular network and WLAN for non-seamless multi-access mode.
- **Enabler 3 – SD-WAN**
  Established Software-Defined Wide Area Network (SD-WAN) solutions are being enhanced with host functionality, which can then be operated in the vehicle system layer. Established SD-WAN policies define traffic steering policies, including whether to route packets directly to the internet or via secured tunnels. These SD-WAN policies can be enhanced to enable traffic steering based on access network characteristics. This solution applies to cellular and WLAN networks for seamless multi-access mode.

In Table 2, the solutions for traffic steering are summarized based on where the traffic steering decision is made, conditions and requirements. The details of each solution are introduced afterward.

Table 2. Summary of solutions for traffic steering.

Solutions	Layers where traffic steering decision is made	Option number indicated in Figure 36	Assumptions and requirements
Solution 4.1 – MPTCP/MPQUIC	At the system layer	2 and 5	Both the vehicle system and mobility service instance support MPTCP/MPQUIC or there is a function to do the mapping between TCP/QUIC and MPTCP/MPQUIC.
Solution 4.2 – Generic Multi-Access (GMA)/Multi-Access Management Services (MAMS)	At the system layer	2 and 5	Both the vehicle system and mobility service instance support GMA/MAMS.
Solution 5.1 – LTE WLAN Aggregation (LWA)	At the access layer	4	Applies only to the LTE network; the cellular network must support LWA.
Solution 5.2 – LTE WLAN radio-level integration over IPSec tunnel (LWIP)	At the access layer	4	Applies only to the LTE network; the cellular network must support LWIP.
Solution 5.3 – Access Traffic Steering, Switch and Splitting (ATSSS)	At the access layer	4	Applies only to the 5G network; the vehicle system and the cellular network must support ATSSS.
Solution 5.4 – Multi Radio Dual Connectivity (MR-DC)	At the access layer	4	Both the vehicle system and the cellular network must support MR-DC.

Solution 4 – System Layer Solutions

Solution 4.1 – MPTCP/MPQUIC

Multipath TCP (MPTCP) is a major modification to TCP that allows multiple paths to be used simultaneously by a single transport connection. The MPTCP protocol has been standardized by the IETF in RFC 6824. MPTCP allows multiple sub-flows to be set up for a single MPTCP session. An MPTCP session starts with an initial sub-flow. Then, after the first MPTCP sub-flow is set up, additional sub-flows can be established that are bound to the existing MPTCP session [11]. Data for the connection can then be sent over any of the active sub-flows that have the capacity to take it.

Quick UDP Internet Connection (QUIC) is a multiplexed and secure transport protocol that runs on top of UDP and combines functions of HTTP/2, TLS and TCP. QUIC is targeted at reducing the latency of client-server communication, providing an alternative to the conventional layered HTTP/TLS/TCP protocol stack used by the web. One of the rationales for the development of QUIC has been the constraints experienced by TCP that is implemented in operating system kernels and middlebox firmware, making significant changes to TCP (e.g., multipath capability) very challenging to deploy [11]. Being based on UDP, QUIC does not suffer from such limitations and hence is able to incorporate new features without having to upgrade legacy systems. Multipath

QUIC (MPQUIC) is an extension to the QUIC protocol that enhances the migration capabilities to enable support of a single connection over multiple paths.

In this solution, the decision is made by the MPTCP/MPQUIC layer, which may be requested by the application to set up multi-paths. This solution applies to both cellular networks and WLANs for non-3GPP seamless mode. If ATSSS is supported in the cellular network, MPTCP can be regarded as a “higher-layer” implementation of ATSSS. This solution requires MPTCP or MPQUIC capability at both the vehicle and mobility service instances.

Solution 4.2 – Generic Multi-Access (GMA)/Multi-Access Management Services (MAMS)

GMA/MAMS is a mechanism to steer traffic over different connections. GMA convergence [12] provides a radio-agnostic IP-layer solution to support seamless traffic splitting, switching and steering over multiple connections/paths in the above examples. It can be configured and managed by over-the-top control messages, such as MAMS [13], and consists of the following two sublayers:

- Convergence sublayer: this layer performs multi-access-specific tasks, such as multi-link (path) aggregation, splitting/reordering, lossless switching/retransmission, fragmentation, concatenation, etc.
- Adaptation sublayer: this layer performs functions to handle tunneling, network layer security and NAT (network address translation).

The convergence sublayer operates on top of the adaptation sublayer in the protocol stack (see Figure 37), and uses a new lightweight trailer-based encapsulation protocol [12] for inserting control information, such as sequence number or timestamp, into each data packet. The adaptation sublayer uses existing protocols such as IPsec, DTLS, UDP or NAT.

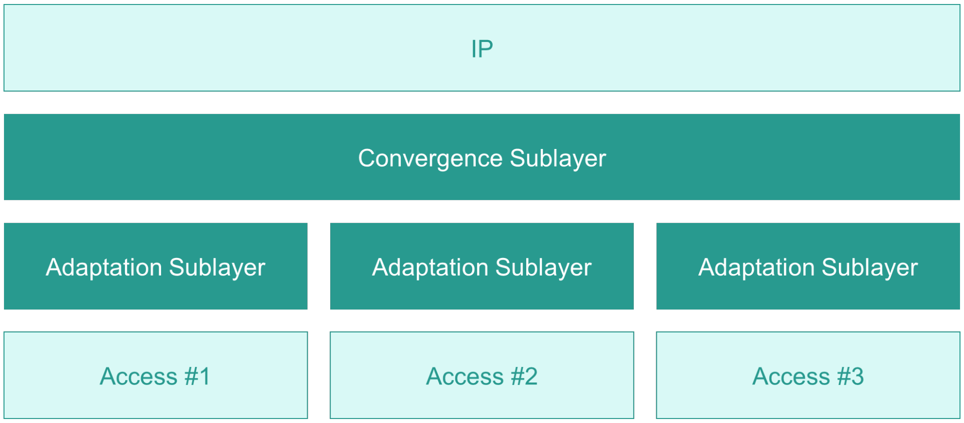

Figure 37. GMA protocol stack.

In this solution, the decision can be made by MAMS negotiation between the Network Connection Manager (NCM) and Client Connection Manager (CCM) or it can be implementation dependent, which further configures the GMA paths. This solution requires that the GMA layer be implemented on both the vehicle and mobility service instances, and it applies to both cellular and WLAN networks for non-3GPP seamless mode.

Solution 5 – Access Layer Solutions

Solution 5.1 – LTE WLAN Aggregation (LWA)

LWA is a 3GPP Release-13 feature to enable LTE and WLAN interworking at RAN level. The LWA radio protocol architecture for the scenarios of co-located eNB and WLAN Termination (WT) is shown in Figure 38 [14].

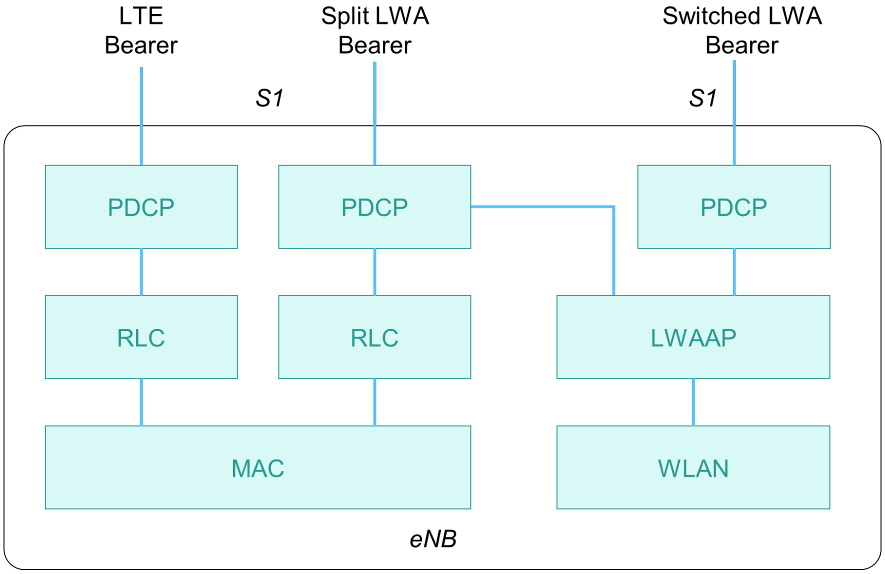

Figure 38. LWA radio protocol architecture stack in a co-located scenario.

A sublayer LTE-WLAN Aggregation Adaptation Protocol (LWAAP) is introduced between a split LWA bearer and the WLAN. Functions of the LWAAP sublayer are performed by LWAAP entities. For an LWAAP entity configured at the eNB, there is a peer LWAAP entity configured at the UE. For all LWA bearers, there is one LWAAP entity in the eNB and one LWAAP entity in the UE. A new interface, Xw, is defined between the eNB and WT in the non-co-located scenario.

In the control plane, the eNB is responsible for LWA activation, deactivation and the decision about which bearers are offloaded to the WLAN. It does so using WLAN measurement information reported by the UE. Once LWA is activated, the eNB configures the UE with a list of WLAN identifiers (referred to as the WLAN mobility set) within which the UE can move without notifying the network.

In the data plane, for PDUs sent over a WLAN in LWA operation, the LWAAP entity, as specified in a 3GPP specification [14], generates an LWAAP PDU containing a Data Radio Bearer (DRB) identity, and the WT uses a defined sequence for forwarding the data to the UE over the WLAN.

In this solution, the decision is made by the eNB to select a split bearer based on RAN-level information such as measurements and the UE's preference. This solution only applies to an LTE network for 3GPP-based seamless mode between LTE and WLAN networks, and it requires an LWAAP entity configured at the vehicle and an eNB at the cellular network.

Solution 5.2 – LTE WLAN radio-level integration over IPSec tunnel (LWIP)

LWIP is also a 3GPP feature to enable LTE and WLAN networks to securely interwork at the RAN level. The LWIP feature allows UE in RRC_CONNECTED to be configured by the eNB to utilize WLAN radio resources via IPsec tunneling [16]. The end-to-end protocol architecture for LWIP is illustrated in Figure 39. Connectivity between the eNB and the LWIP-SeGW is provided by the Xw interface introduced in Solution 5.1.

The IP Packets transferred between the UE and the LWIP-SeGW are encapsulated using IPsec, as specified in a 3GPP specification [16], in order to provide security to the packets that traverse the WLAN. The IP packets are then transported between the LWIP-SeGW and eNB via the Xw interface. The end-to-end path between the UE and eNB via the WLAN network is referred to as the LWIP tunnel.

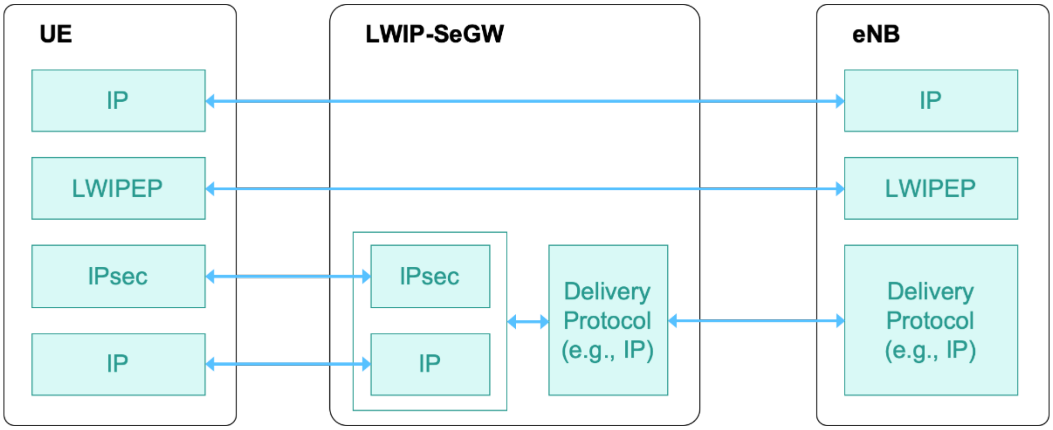

Figure 39. Protocol architecture for LWIP.

A single IPSec tunnel is used per instance of UE for all the data bearers that are configured to send and/or receive data over the WLAN. The data corresponding to each IPSec tunnel is transported over the Xw interface on a single GTP-U tunnel. Each data bearer may be configured so that traffic for that bearer can be routed over the IPsec tunnel in downlink only, uplink only or both uplink and downlink over the WLAN. SRBs are carried over the LTE network only. The eNB configures specific bearers to use the IPsec tunnel.

For the downlink of a data bearer, the packets received from the IPsec tunnel are forwarded directly to upper layers. For the UL, the eNB configures the UE to route the uplink data either via LTE or WLAN networks using Radio Resource Control (RRC) signaling. If routed via the WLAN, then all UL traffic of the data bearer is offloaded to the WLAN.

In this solution, the decision is made by the eNB to select a split bearer based on RAN-level information such as measurements and the UE's preference. This solution requires LWIPEP entities configured at both the UE and the eNB, and it requires the WLAN AP to be enhanced to support the 3GPP-defined Xw interface. This solution only applies to LTE networks for 3GPP-based seamless mode between LTE and WLAN networks.

Solution 5.3 – Access Traffic Steering, Switch and Splitting (ATSSS)

ATSSS provides a way to manage 3GPP access and non-3GPP access such as WLANs, and to manage the traffic from both networks when the UE alternates between these networks, as defined in a 3GPP specification [17]. The reference architecture is shown in Figure 40 [17]. The WLAN connects to the cellular core network by the Non-3GPP Inter-Working Function (N3IWF).

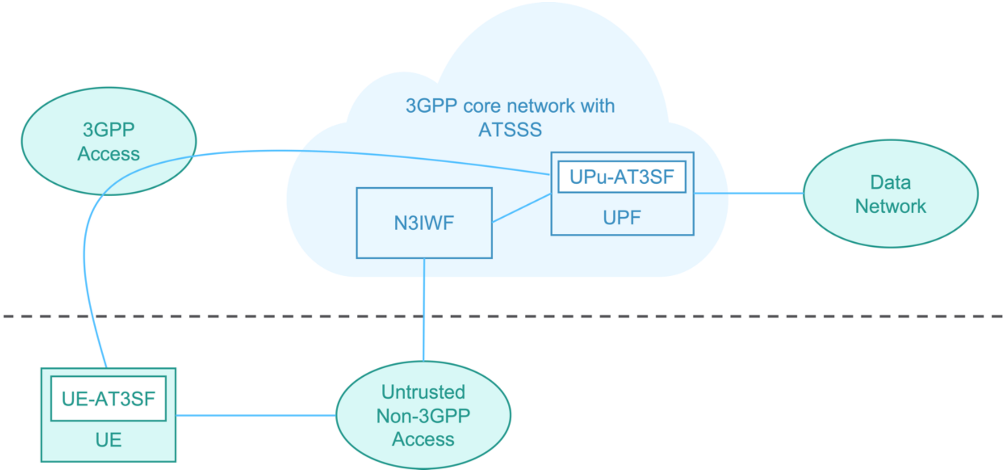

Figure 40. ATSSS reference architecture.

In ATSSS, a Multi-Access PDU (MA-PDU) session is created by bundling together two separate PDU sessions, which are established over different accesses. ATSSS policies are provisioned in the UE and the UPF, which can be generated in the PCF to provide the access switching rule and put one access in stand-by mode.

Within the same MA-PDU session, if MPTCP is enabled, it is possible to steer the MPTCP flows by using the MPTCP protocol (or the MPTCP function) and, simultaneously, to steer all other flows by using lower-layer steering functionality, called the ATSSS function. This is schematically illustrated in Figure 41 [17] for the UE. Note that the same set of ATSSS rules is applied to configure the MPTCP function and the ATSSS function. During the process of setting up the MA-PDU session, the UE shall exchange capability information about support of MPTCP with the core network as specified in a 3GPP specification [17].

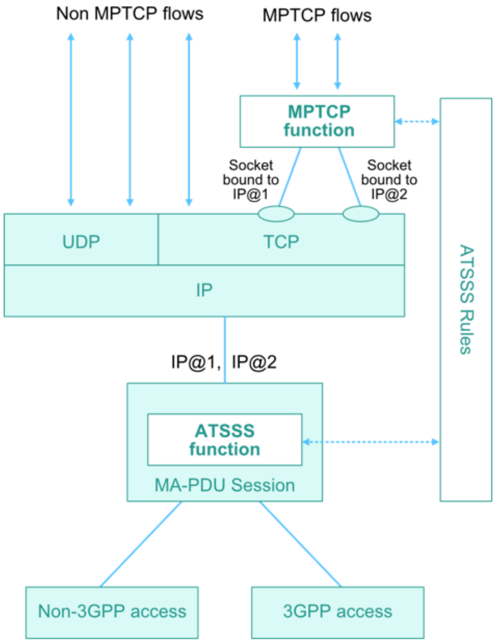

Figure 41. Example of UE supporting an MPTCP function and an ATSSS function.

In this solution, the UE or UPF can initiate the MA-PDU session to trigger ATSSS, and the traffic is managed by the ATSSS policies provided by the PCF. ATSSS is a solution to enable 3GPP-based seamless traffic steering over a cellular and a WLAN network. This solution applies to 5G networks and requires the UE and cellular core network to support ATSSS capabilities. MPTCP can be regarded as a higher-layer implementation of ATSSS.

Solution 5.4 – Multi Radio Dual Connectivity (MR-DC)

MR-DC is an extension of dual connectivity that allows UE to simultaneously connect to two nodes in the network [18]. In MR-DC, multiple Rx/Tx UE may be configured to utilize resources provided by two different nodes connected by non-ideal backhaul. One node acts as the Master Node (MN) and the other as the Secondary Node (SN). The MN and SN are connected by a network interface and at least the MN is connected to the core network. The MN may control the connectivity and data splitting toward the two nodes. Note that "non-ideal backhaul" generally means backhauls with latency of 5~30 ms or even higher.

MR-DC can generally increase user throughput, provide mobility robustness and support load balancing among eNBs/gNBs.

In this solution, the decision about where to steer the traffic is made by the MN based on information such as measurements. This solution requires MR-DC capability at the vehicle and the cellular network, and it applies to both LTE and 5G networks for 3GPP-based seamless mode.

3.4.3 Conclusions

We recommend the following combination of solutions for access network selection.

Connection selection

The AECC recommends taking the decision about connection selection in the system layer of the vehicle system, as described in Solution 1.2, where the connection algorithm considers information such as signal strength, cost and policies. Applications running in the vehicle system can provide policies to the connection selection function. The connection selection function then interacts with communication modules in the access layer to connect to and disconnect from available access networks. Solution 1.2 can reduce the complexity of the vehicle system by a unified interface to applications to avoid duplications of access network selection assistance functions at the application layer.

Traffic steering

When multiple connections are available and used, the vehicle system steers traffic on the transport layer using MPTCP/MPQUIC, as described in Solution 4.1. This allows decoupling the traffic steering task from the access networks that are used. Consequently, this solution works in the same way for different combinations of access networks, such as using multiple cellular networks, or a combination of a WLAN and a cellular network.

3.5 Provisioning and Configuration Update

3.5.1 Key Issue

The AECC system embraces high-volume data loads of different varieties. As the vehicle system moves, the location, environment, network availability and generated data change dramatically. During the data exchange process among the vehicle system, access networks and mobility service instances, a great many parameters and policies are involved. Some parameters and policies can be fairly static, but some can be very dynamic. In order to prepare mobility service instances, access networks and the vehicle systems to meet AECC system service requirements in a dynamic environment, provisioning and configuration updates need to be supported as shown in Figure 42. The AECC system should be aware of these changes either based on access network or vehicle reports, and adjust the policy/configuration accordingly. The cellular network can also update its related parameters and policies to the vehicle system. In this section, the parameters and policies related to the AECC system are characterized and possible solutions are summarized.

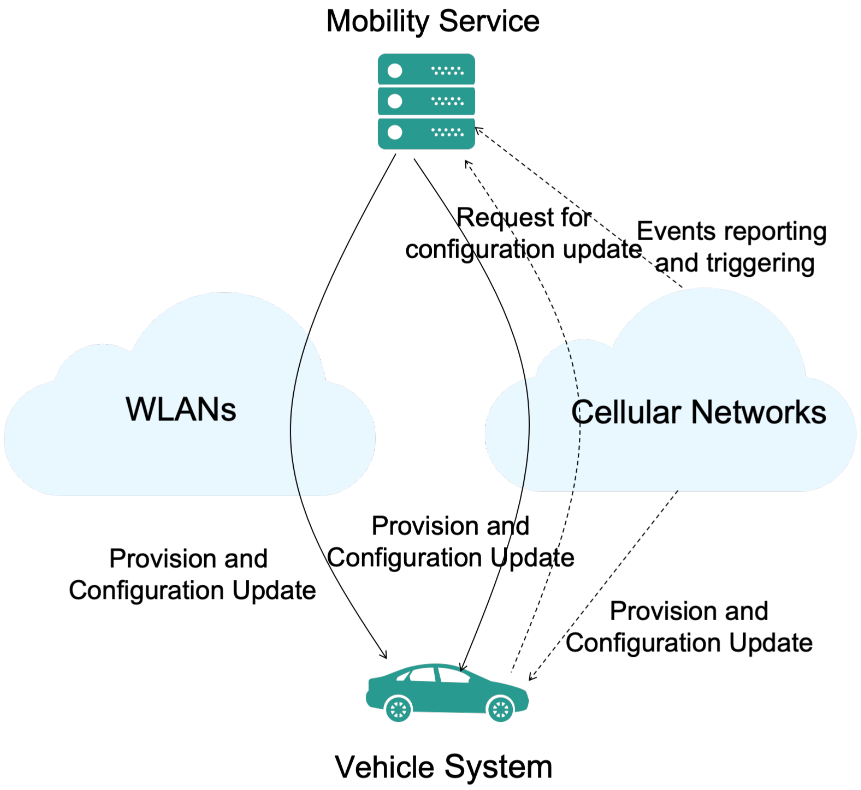
Mobility Service
Request for configuration update
Events reporting and triggering
WLANs
Cellular Networks
Provision and Configuration Update
Provision and Configuration Update
Provision and Configuration Update
Vehicle System

Figure 42. Provisioning and configuration update.

Provisioning and configuration for the vehicle system mainly include provisioning parameters and policies; for example:

- Credentials and configurations of access networks, e.g., Access Point Name (APN) and Quality of Service (QoS)
- Non-access-network-related configurations, e.g., hostnames and addresses
- Policies, e.g., edge offloading policies

3.5.2 Potential Solutions

Solutions for provisioning and configuration updates are listed below.

- Solution 1 – Preconfiguration
- Solution 2 – Configured through the Access Network
    - Solution 2.1 – Configured by cellular subscription
    - Solution 2.2 – Configured through the cellular network
- Solution 3 – Provisioned by the AECC System Server at the Application Level
- Solution 4 – Provision through a Generic AECC System Configuration Function
- Solution 5 – Provision a Bootstrap URL of a Configuration Function
    - Solution 5.1 – Bootstrap URL provisioned by the DHCP server
    - Solution 5.2 – Preconfigured bootstrap URL

In Table 3, the solutions for provisioning and configuration update to the vehicle are summarized, with applicable examples of parameters and polices. Note that this is not an exhaustive list but is instead examples.

Table 3. Solutions for provision and configuration update.

Solutions	**Examples of parameters and policies**	**Static or dynamic**
Solution 1 – Preconfiguration	Credentials for cellular or WLAN networks from cellular or WLAN MNOs; IP addresses or domain names, URLs of potential mobility service instances, non-AECC servers from the AECC system; user preference from the user; APNs and service QoS from the access network and MSP	Generally static
Solution 2.1 – Configured by cellular subscription	Cellular subscription information such as device category, billing information, parameters to access the network, data rate, data transmission window, regional regulatory requirements agreed between MNOs and MSPs or manufacturers	Generally static
Solution 2.2 – Configured through the cellular network	Update UE-related information such as the UE's behavior and its communication groups; negotiate a policy, such as a background data transfer policy; influence the traffic, such as data offloading, etc.	Generally dynamic
Solution 3 – Provisioned by the AECC system application server at the application level	Application-level parameters on the vehicle system from applications on the AECC system servers	Both static and dynamic; out of the scope of the AECC

Solutions	Examples of parameters and policies	Static or dynamic
Solution 4 – Provision through the AECC system configuration function	AECC system-related parameters such as IP addresses or domain names, URLs of potential mobility service instances, non-AECC system servers; can integrate user preference.	Both static and dynamic
Solution 5.1 – Bootstrap URL provisioned by the DHCP server	Appropriate parameters and policies in a configuration file sent together with the IP configurations when the vehicle first accesses the network	Both static and dynamic
Solution 5.2 – Preconfigured bootstrap URL	Appropriate parameters and policies in a configuration file sent as a bootstrap URL to the vehicle	Both static and dynamic

3.5.2.1 Solution 1 – Preconfiguration

The vehicle can be preconfigured (e.g., by its manufacturer) with the policies for the AECC system, which may include:

- Credentials for cellular or WLAN network domains, agnostic about the AECC system, such as SIM card or WLAN account information. This provision information is expected to come from cellular or WLAN operators.
- Configurations for mobility services such as IP addresses or domain names, URLs of potential mobility service instances, non-AECC system servers, agnostic about the cellular or WLAN network. This provision information is expected to come from the AECC system.
- Preferences at the application level that may have an impact on mobility services, such as user settings, account and billing information or data transmission preference. This information is expected to come from the user.
- Configuration for the cellular or WLAN network to fulfill mobility services such as APN and service QoS. This information is expected to come from the access network and MSP.

This solution includes the information that shall be preconfigured in the application or platform of the vehicle system. This solution applies to both cellular and WLAN access networks.

3.5.2.2 Solution 2 – Configured through the Access Network

The solutions described in this section mostly target policies relating to cellular network communication.

Solution 2.1 – Configured by cellular subscription

Relatively static parameters and policies can be provisioned to the vehicle system as part of the cellular network subscription information as agreed between the network operator and vehicle manufacturer (or MSP). This may include device category, billing information, parameters to access the network, data rate, data transmission window, regional regulatory requirements and so on. Subscription information is provisioned by the access network, which is different from Solution 1.

This solution applies to the cellular access network only.

Solution 2.2 – Configured through the cellular network

In one alternative version of this solution, the mobility service instance provisions the cellular network via the SCEF/NEF over the T8 reference point or Nnef interface. The cellular network can then enforce the vehicle system's policy via the UE's configuration update procedure as defined in a 3GPP specification [22]. The parameters and policies can be updated by the mobility service instances on a periodic, event-triggered basis. This applies to the access network-related parameters and policies.

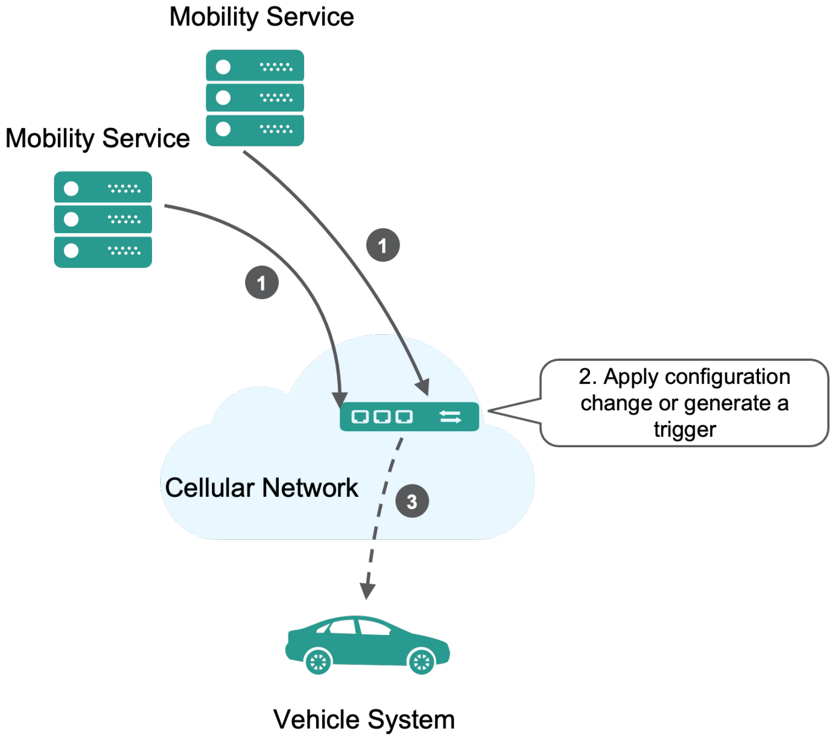

Figure 43. Procedure for provision configuration via cellular network.

The parameter and policy update enforced by the cellular network is shown in Figure 43. The procedure is described as follows:

1) The MSP center server or edge server sends configuration parameters through APIs to the SCEF/NEF in the cellular network via T8 or Nnef.
2) The SCEF/NEF applies configuration changes to cellular network entities, which may include a PCRF/PCF, HSS/UDM, UDR, etc. As stated in clause 4.15.6 of a 3GPP specification [22], the mobility service instances can leverage an AF to externally provision parameters to UDM/UDR.
3) The updated parameters are delivered to the vehicle system via N1/S1 or Uu. This step is optional: that is, the vehicle system does not need to be specifically configured; the policies will be applied in the network when the UE performs access or mobility-related procedures.

The SCEF/NEF can expose capabilities for mobility service instances to

- Update UE-related information such as the UE's behavior and its communication groups.
- Negotiate a policy, such as a background data transfer policy.

- Influence the traffic, such as data offloading, etc.

More policies are defined in a 3GPP specification [22]. Figure 44 shows the procedure for an AF requesting traffic influence on the UE by modifying session-related policies in the PCF.

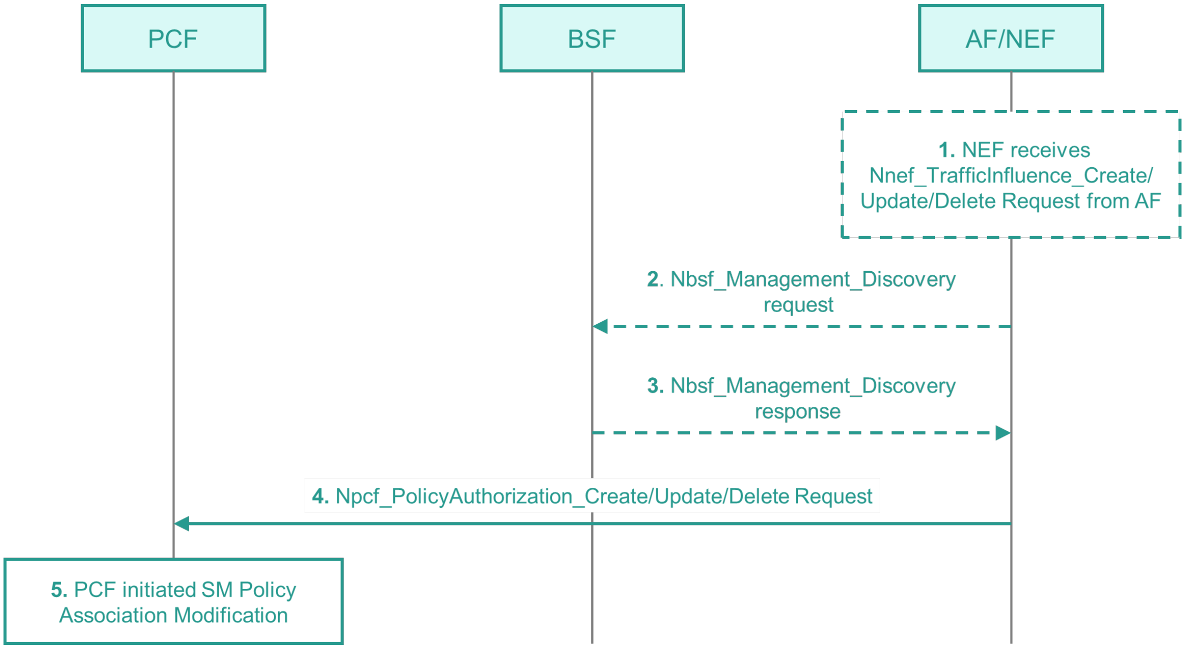
PCF
BSF
AF/NEF
1. NEF receives Nnef_TrafficInfluence_Create/ Update/Delete Request from AF
2. Nbsf_Management_Discovery request
3. Nbsf_Management_Discovery response
4. Npcf_PolicyAuthorization_Create/Update/Delete Request
5. PCF initiated SM Policy Association Modification

Figure 44. AF requesting a configuration/policy update.

1) The AF requests to influence the UE’s traffic routing through the SCEF/NEF by the UE’s address.
2) [Optional] The SCEF/NEF initiates the discovery of a related PCF by the UE’s address.
3) [Optional] The Bootstrapping Server Function (BSF) provides the PCF’s address in the discovery response.
4) The AF/NEF initiates the policy authorization to create/update/delete the requested policy.
5) The PCF starts to create/update/delete the related policies.

In addition, the PCF can update UE access selection and PDU session selection-related policy information in the UE configuration, as shown in Figure 45.

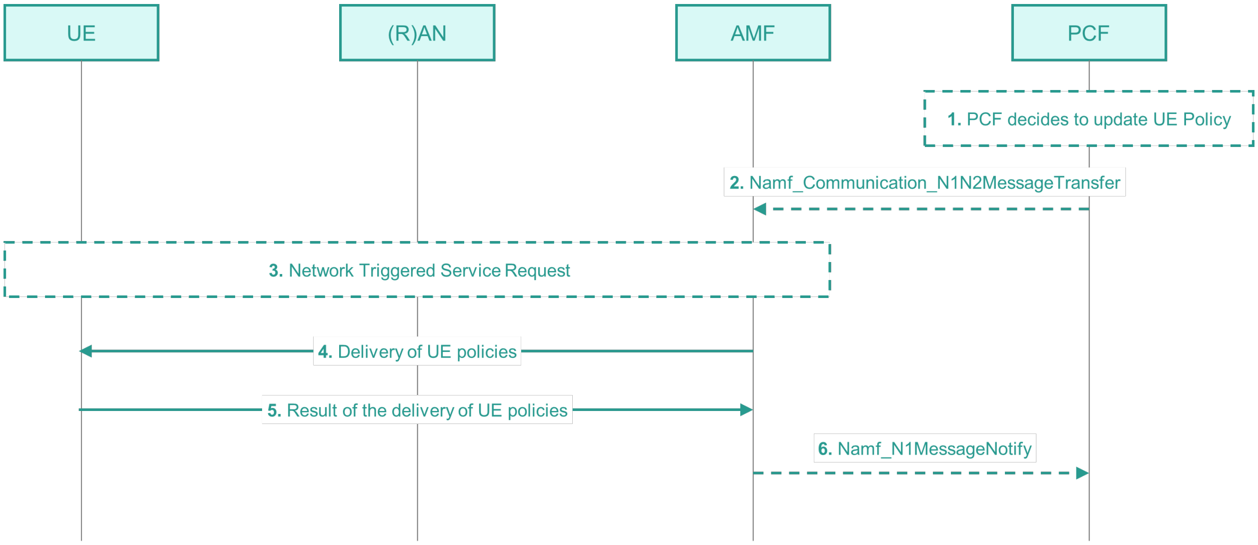
UE
(R)AN
AMF
PCF
1. PCF decides to update UE Policy
2. Namf_Communication_N1N2MessageTransfer
3. Network Triggered Service Request
4. Delivery of UE policies
5. Result of the delivery of UE policies
6. Namf_N1MessageNotify

Figure 45. UE configuration update procedure.

1) The PCF decides to update UE policy procedures based on triggering conditions.
2) The PCF invokes the Namf_Communication_N1N2MessageTransfer service operation provided by the AMF.
3) [Optional] The UE may need to be triggered to initiate a service request procedure for policy delivery.
4) The AMF delivers related policies to the UE.
5) The UE confirms the policy delivery.
6) [Optional] The AMF notifies the PCF about the policy delivery.

In another alternative version of this solution, the mobility service instance requests an SMS trigger from the cellular network that includes the vehicle system-related parameters. Upon receiving the MSP request (via the SCEF/NEF), the cellular network will generate the SMS and send it to the vehicle system. The cellular network does not need to understand the configuration update. The procedure is described as follows.

1) The mobility service instances request a trigger from the cellular network that includes the vehicle system-related parameters. The MSP request is sent through the NEF.
2) The cellular network generates an SMS with server-configured parameters.
3) The cellular network sends the SMS message to the vehicle system.

Such a solution can be used to provision small configuration files to vehicle systems, for example.

3.5.2.3 Solution 3 – Provisioned by the AECC System Server at the Application Level

At the application level, the parameter and policy provisioning is done through the interface between the applications residing on the vehicle system (e.g., an AECC platform service or a third-party application hosted by the AECC system) and the applications residing on mobility service instances. If the applications running on the vehicle system side and on the mobility service instance side are third-party applications, the technical realization of the application layer interface is outside the scope of the AECC.

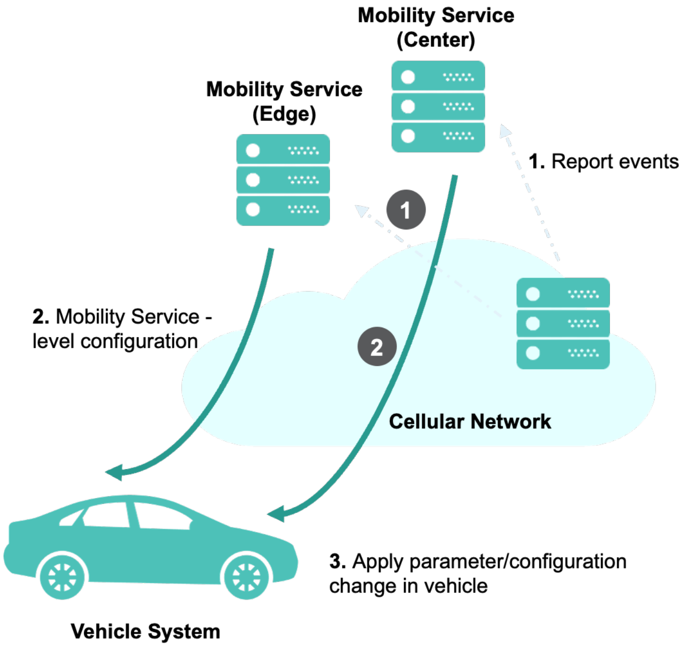
Mobility Service (Center)
Mobility Service (Edge)
1. Report events
1
2. Mobility Service - level configuration
2
Cellular Network
3. Apply parameter/configuration change in vehicle
Vehicle System

Figure 46. Application-level parameter/configuration update.

Figure 46 shows the application-level parameter and policy update.

1) The cellular network optionally reports network events to mobility service instances. The network events could have been configured previously through the NEF. When using other types of access network, entities within those access networks may also provide reports to mobility service instances.
2) The mobility service instances update parameters and configuration through the AECC application, which is agnostic about the change of access network. Alternatively, the mobility service instances can send a push notification with provisioned parameters, which has been discussed in Section 3.3.1, the Key Issue in vehicle system reachability.
3) The vehicle system applies the related parameter change.

The monitoring events can include loss of connectivity, location reporting, number of instances of UE presented in a geographical area, etc. as defined in Section 4.15.3 in a 3GPP specification [22].

3.5.2.4 Solution 4 – Provision through a Generic AECC System Configuration Function

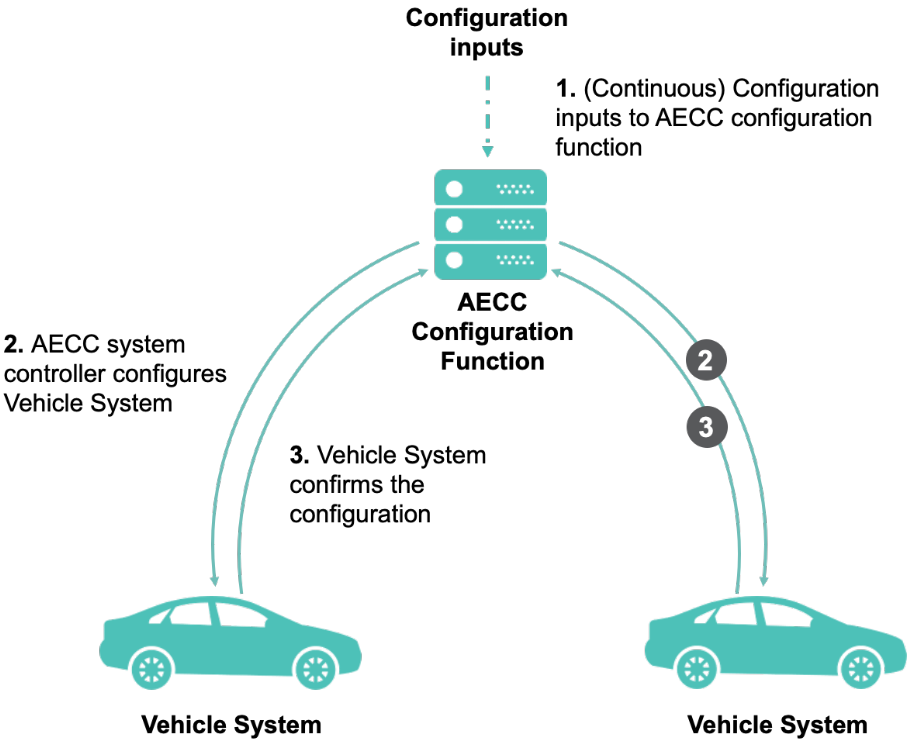
Configuration inputs
1. (Continuous) Configuration inputs to AECC configuration function
AECC Configuration Function
2. AECC system controller configures Vehicle System
3. Vehicle System confirms the configuration
2
3
Vehicle System
Vehicle System

Figure 47. Configuration by the AECC system configuration function (AECC system configuration function-triggered).

The vehicle system is configured through the AECC system configuration function if there is an interface between the configuration function and the vehicle system. In one alternative version of the solution, the vehicle system configuration is triggered by the AECC system configuration function, as shown in Figure 47. The steps are as follows:

1. [Optional] Configuration parameters are input into the AECC system configuration function. The inputs can be provided by external users or applications through interfaces provided by the AECC system configuration function. This step can be event-triggered or periodic.
2. The AECC system configuration function provides provision and configuration updates to the vehicle system.
3. The vehicle system confirms to the AECC system configuration function when the configuration updates are received.

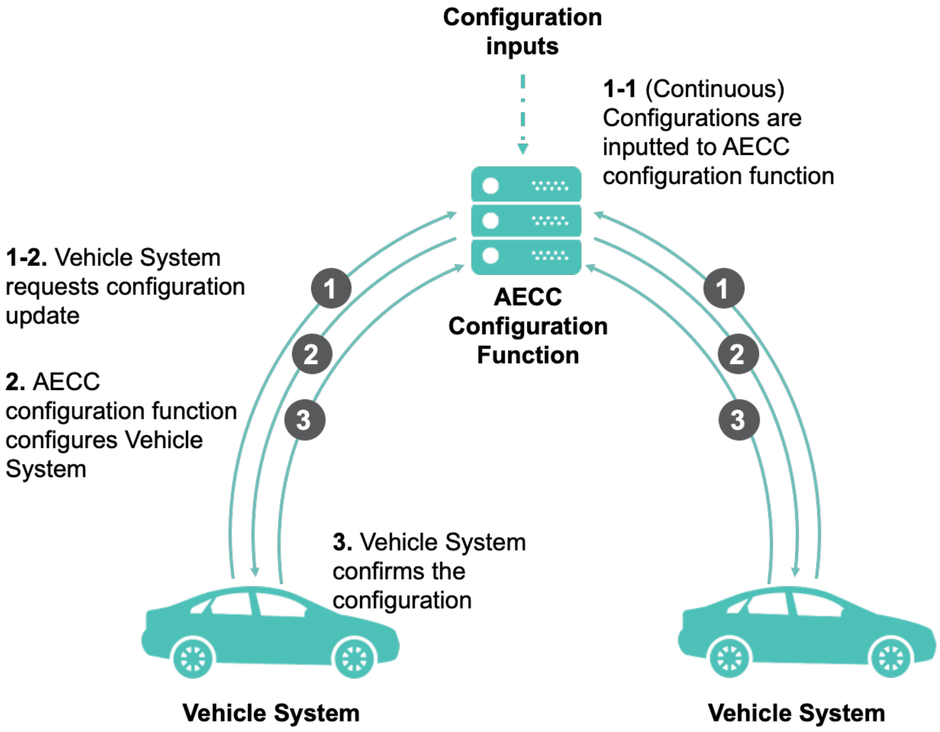
Configuration inputs
1-1 (Continuous) Configurations are inputted to AECC configuration function
AECC Configuration Function
1-2. Vehicle System requests configuration update
2. AECC configuration function configures Vehicle System
3. Vehicle System confirms the configuration
1
2
3
1
2
3
Vehicle System
Vehicle System

Figure 48. Configuration by the AECC system configuration function (vehicle system-triggered).

In another alternative version of the solution, the vehicle system configuration is triggered by the configuration function, as shown in Figure 48. The steps are as follows:

1-1) [Optional] Configuration parameters are input into the AECC system configuration function. The inputs can be provided by external users or applications through interfaces provided by the AECC system configuration function. This step can be event-triggered or periodic.

1-2) The vehicle system sends a configuration update request to the AECC system configuration function. Note that the order of sequence between Step 1-1 and Step 1-2 is interchangeable.

2) The AECC system configuration function provides provision and configuration updates to the vehicle system.

3) The vehicle system confirms to the AECC system configuration function when the configuration updates are received.

3.5.2.5 Solution 5 – Provision a Bootstrap URL of a Configuration Function

This solution includes a two-step provisioning mechanism, where the first step targets provisioning a URL (bootstrap URL) from which a more extensive configuration file is downloaded in the second step. The configuration server can be outside the AECC system.

Solution 5.1 – Bootstrap URL provisioned by the DHCP server

In this solution, the location of the configuration file is sent together with the IP configurations when the vehicle system first accesses the network. This scheme works for configuration that applies directly to the vehicle system upon the vehicle system's startup. The DHCP bootstrap option is described in IETF RFC 2132 [19] with additional information in RFC 5970 for DHCPv6 [20]. One condition for applying this solution is that the DHCP server needs to be informed of the configuration function location. If the DHCP server and the server hosting the configuration function are provided by different service providers, inter-service provider agreement is needed.

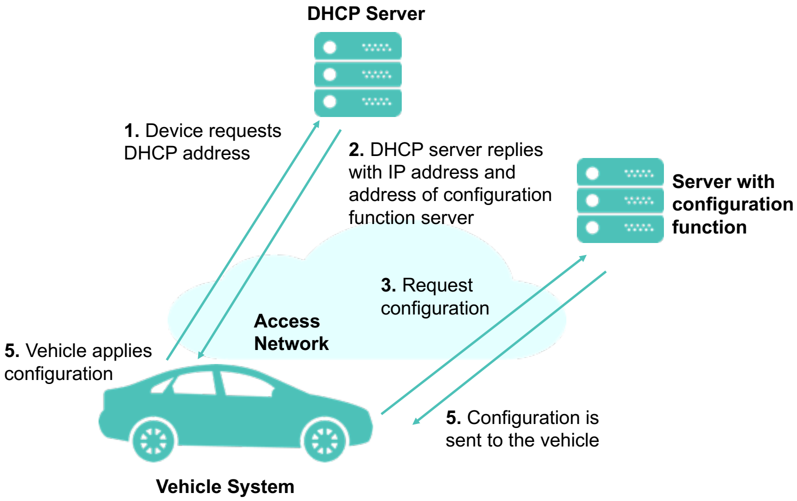
DHCP Server
1. Device requests DHCP address
2. DHCP server replies with IP address and address of configuration function server
Server with configuration function
3. Request configuration
Access Network
5. Vehicle applies configuration
5. Configuration is sent to the vehicle
Vehicle System

Figure 49. Configuration file distribution through the DHCP.

Figure 49 shows configuration file distribution using the DHCP.

1) The vehicle system requests the IP address from the DHCP server.
2) The DHCP server sends IP configurations along with a configuration function location.
3) The vehicle system accesses the configuration file from the location provided previously by the DHCP server.
4) The configuration file server sends the configuration file to the vehicle system.
5) The vehicle system applies the configurations.

Solution 5.2 – Preconfigured bootstrap URL

In this scheme, the location of the configuration file server is preconfigured in the vehicle system and the configuration files are downloaded when there is connectivity to the configuration function. Typically, the URL would contain a hostname that needs to be resolved using DNS, before contacting the configuration function.

Here, “preconfigured” does not necessarily mean permanently hard-coded, but that the bootstrap URL is always available in the vehicle system, possibly by having an initial hard-coded URL, which can be updated periodically during the lifetime of the vehicle system using other provisioning mechanisms.

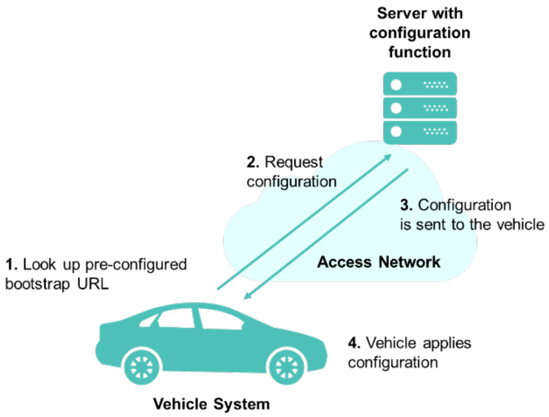
Server with configuration function
2. Request configuration
3. Configuration is sent to the vehicle
Access Network
1. Look up pre-configured bootstrap URL
4. Vehicle applies configuration
Vehicle System

Figure 50. Configuration file at preconfigured location.

Figure 50 shows the configuration look-up procedure.

1) The vehicle system looks up the bootstrap URL.
2) The vehicle system requests the configuration file from the preconfigured location.
3) The configuration file server sends the configuration file to the vehicle system.
4) The vehicle system applies the configuration.

While the described procedure assumes a vehicle system as target for the provisioning, the solution is also applicable for other entities, such as mobility service instances.

This solution applies to both cellular and WLAN Access Networks.

3.5.3 Conclusions

We recommend the following solutions for the respective layers:

- For provisioning and configuration updates in the system layer, the AECC recommends using Solution 1 (Preconfiguration) for static parameters and policies and for initialization. Solution 5.2 (Preconfigured bootstrap URL) is recommended for semi-static provisioning and configuration updates, possibly updating preconfigured parameters. As a more advanced method, Solution 4, Provision through a Generic AECC System Configuration Function, can be used for semi-static and dynamic provisioning and configuration updates, given that corresponding interfaces are available and configured.
    - One example of provisioning in the system layer would be provisioning a mobility service instance FQDN to a vehicle system, which might be preconfigured initially (Solution 1), with the option to update the FQDN later using Solution 4 and/or Solution 5.2.
- For provisioning and configuration updates in the access network layer, in the case of cellular networks, we recommend Solution 2.1 (Configured by cellular subscription), and Solution 2.2 (Configured through the cellular network), depending on the parameters. The condition for applying Solution 2.2 is that the T8 (in EPS) or Nnef/N33 (in 5GS) interfaces and related APIs may be needed in the access layer. Specific mechanisms for provisioning concrete parameters regarding the cellular network are specified by 3GPP. For any access network type, Solution 1 (Preconfiguration) is recommended for static parameters, and for default values.
    - APNs and DNNs are preconfigured in the vehicle system (Solution 1).

  - UE categories are configured as part of the cellular subscription (Solution 2.1).

Provisioning in the application layer (i.e., from the mobility service instance application layer to the vehicle system application layer) is outside the scope of the AECC.

3.6 Opportunistic Data Transfer

3.6.1 Key Issue

Network capacity planning has become a major challenge, especially for MNOs, due to the exponential increase in mobile data traffic. Meanwhile, new vertical markets such as connected vehicles will further drive mobile traffic growth to a new level. It causes a critical need for MNOs and MSPs to provide sufficient network capacity to meet the traffic demands of new mobility services.

On the other hand, data traffic of non-latency-sensitive mobility services, such as vehicle data collection [5], could be transferred opportunistically, which implies that it will be delivered in a lower priority without affecting other services delivered by the network. The goal of opportunistic data transfer is to deliver certain types of delay-tolerant data, such as vehicular data collection, without degrading the service quality of other data traffic.

Note 1: due to the different dynamics of mobile data traffic in different timescales, opportunistic data transfer could be managed differently. For instance, mobile data traffic usually fluctuates drastically in small-scale time periods (such as a second), while it exhibits a comparatively static pattern in large-scale time (such as over a day or week). Different solutions can be used and combined to make use of these two effects.

Note 2: opportunistic data transfer shall also apply to other access networks defined in the AECC system, such as WLANs. However, as the capacity issues are more urgent to solve with respect to cellular networks, this key issue focuses on the cellular network.

3.6.2 Potential Solutions

The following solutions are described for this key issue.

- Solution 1 – Access Control and Barring
  - Solution 1.1 – Application-specific Congestion control for Data Communication (ACDC)
  - Solution 1.2 – Unattended Data Traffic (UDT)
- Solution 2 – Background Data Transfer (BDT)
  - Solution 2.1 – 3GPP network-based Background Data Transfer (BDT)
  - Solution 2.2 – 3GPP UE-based BDT
- Solution 3 – Dynamic Policy Adaptation

3.6.2.1 Solution 1 – Access Control and Barring

Access control methods are generally considered in order to block or defer certain user groups/services at base stations according to the access class pre-assigned to different users and services (e.g., data and voice) when the network is overloaded. The basic operation is that the UE will block or defer the traffic transmission request according to the broadcast access control policy and parameters from base stations. There are many access control mechanisms that are defined in 3GPP specifications [21], including Access Class Barring (ACB), Service Specific Access Control (SSAC), Extended Access Barring (EAB), Smart Congestion Mitigation (SCM), Application-specific Congestion control for Data Communication (ACDC) and Unattended Data Traffic (UDT). By applying these mechanisms only to the non-latency-sensitive traffic, disturbing other communication is avoided, and opportunistic data transfer is achieved.

Solution 1.1 – Application-specific Congestion control for Data Communication (ACDC)

In 3GPP Release 13, ACDC is standardized, and can be applied to solve the opportunistic data transfer problem. ACDC is an access control mechanism for the operator to allow/prevent new access attempts from particular operator-identified applications in the UE in idle mode. ACDC does not apply to the UE in connected mode. In ACDC, the network can handle access control policy per application category, thus preventing network congestion due to the large amount of delay-tolerant data transfer from vehicles.

Solution 1.2 – Unattended Data Traffic (UDT)

UDT is defined in 3GPP as data traffic that the user is unaware that he/she initiated; for example, based on the screen/keypad lock being activated, length of time since the UE last received any input from the user, or specific types of applications. The IE (Information Element) *SystemInformationBlockType2* contains radio resource configuration information that is common for all UE. The UE can restrict access attempts due to unattended data traffic via the radio resource configuration.

3.6.2.2 Solution 2 – Background Data Transfer (BDT)

The cellular network and mobility service instances can negotiate a time slot (e.g., non-busy hours) to deliver non-time-critical data via the method of background data transfer. It is a proactive approach to solving this key issue, which needs advanced functions of the cellular network, such as APIs for scheduled transmission. Such methods would typically be used for bulk data distribution or upload.

Solution 2.1 – 3GPP network-based Background Data Transfer (BDT)

3GPP specifies that a northbound interface is between the core network entities and third-party application servers using the Service Capability Exposure Function (SCEF, EPC) and Network Exposure Function (NEF, 5GC). This interface specifies APIs that allow a third-party application server to access the network service and capability provided by 3GPP network entities. It is named as the T8 reference point in LTE. In a 5GS, the NEF reuses most of the T8 APIs. An AF could negotiate a BDT policy through the SCEF/NEF and then related policies can be generated in the PCF to be enforced by other network functions, such as AMF for access control or SMF for session management, as shown in Figure 51.

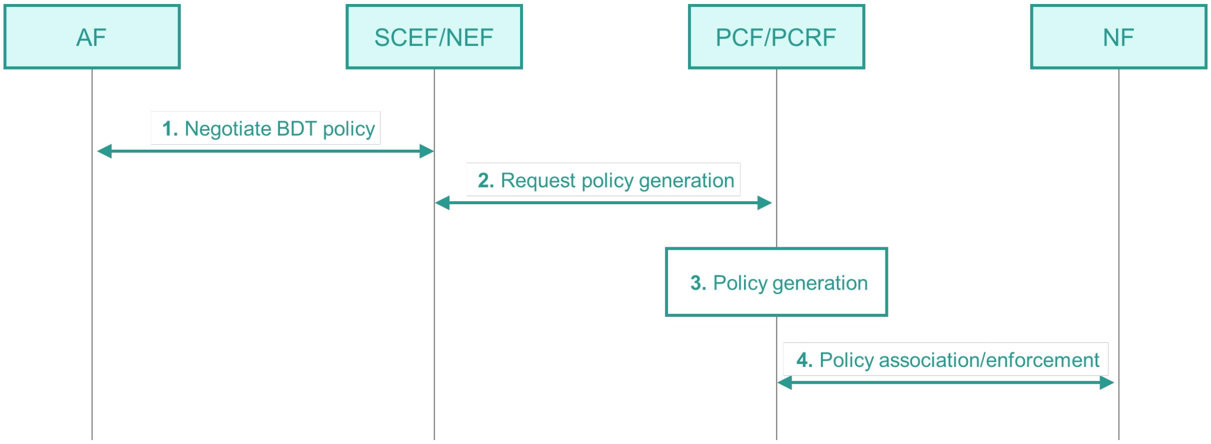

Figure 51. Negotiation for future background data transfer in cellular networks.

The workflow is as follows:

- Step 1: The AF requests to negotiate BDT policy with the SCEF/NEF based on different parameters.
- Step 2: The SCEF/NEF requests policy creation from the PCF/PCRF.
- Step 3: BDT policy generation occurs in the PCF/PCRF.
- Step 4: Policies are applied to related cellular network functions.

Opportunistic data transfer policy can be based on BDT policy. A transfer policy consists of a recommended time window for the background data transfer, a reference to a charging rate for this time window and optionally a maximum aggregated bitrate.

The start/stop can also be assisted by SCEF/NEF event monitoring, as shown in Figure 52.

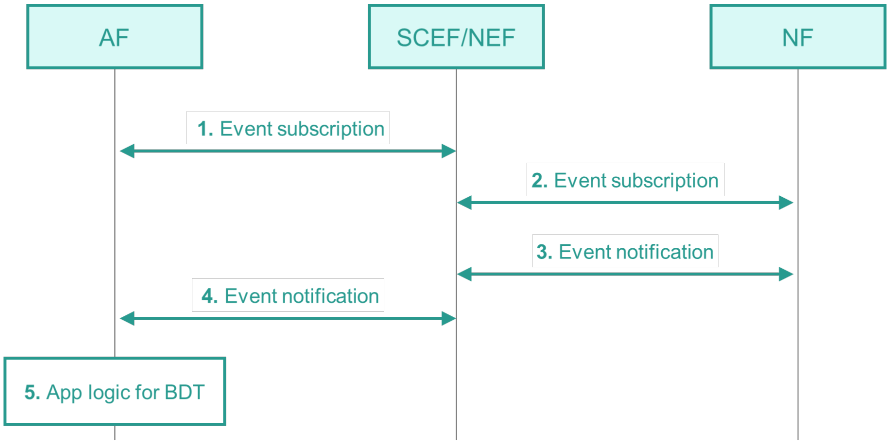

Figure 52. Cellular event triggering for dynamic BDT.

The workflow is as follows:

- The AF requests event monitoring capability from the SCEF/NEF for a vehicle or a vehicle group.
- The SCEF/NEF subscribes to the vehicle's network status.
- A network event is detected at the NF and notified to the SCEF/NEF.

- The AF is notified of the event.
- The AF applies application-related logic for BDT.

The network events can include:

- Vehicle loss of connection
- Vehicle's location
- Network load

For example, the ReportingNetworkStatus API can expose the network status to the mobility service instance, and thus the mobility service instance can manage the data transfer more efficiently. The congestion value in the API indicates whether there is congestion and what the congestion level is. The mechanism can be: first, the mobility service instance subscribes to the network status report according to periodic mode or based on events; second, the mobility service instance receives the congestion indicator from the network status report via the cellular northbound APIs. Finally, according to the congestion indicator, the mobility service instance can decline or continue data transfer to the vehicle system.

Solution 2.2 – 3GPP UE-based BDT

UE can support the 3GPP system to optimize its use of the wireless resources based on policies (such as time window, network area information, etc.) for background data transfer if the policies are sent from the core network or the servers lying outside the 3GPP system.

This solution enables the 5GS to support delivery of the policies for background data transfer to the UE. In particular, the 5G network can provide policies for background data transfer to the UE, so that the 5GS can optimally use the control plane and/or user plane resources by directly managing the transmission behavior of the UE. If background data transfer policy information (such as time window and location criteria) is not going to be sent to the UE as part of the URSP rule, then, at the time the background data transfer is about to start, the AF provides for each instance of UE the background data transfer reference ID together with the AF session information to the PCF (via the N5 interface). The PCF retrieves the corresponding background transfer policy from the UDR and derives the PCC rules for the background data transfer according to this transfer policy.

The policy information content will define the time window and location criteria that need to be met for background data transfer. It defines how and when the PCF activates/distributes the policies related to background data transfer to the UE. A single dedicated PDU session might be used for background data transfer that is established and released based on the background data transfer policies. In addition, the policy should consider avoiding a large quantity of UE sending data and/or signaling to the network at the same time.

In order to ensure that data is sent only in the designated time window, an additional (implementation-dependent) interface between modem and applications is needed in the vehicle system, to deny delivery requests outside this time window. Further, the URSP rules need to be configured in such a way that delivery attempts outside the designated time window are not sent on different PDU sessions by default (e.g., due to "match all" URSP rules).

3.6.2.3 Solution 3 – Dynamic Policy Adaptation

This solution assures that data is received on time but at the same time exploits the relaxed delay constraints of non-real-time communication. Besides describing the data communication aspect, this solution also presents means for the application and network to negotiate background data transmission policies according to a subscription agreement.

It is unlikely that all MNOs will support the same policy enforcement mechanism. The choice of enforcement might have an impact on the client implementation in the vehicles. As a result, harmonizing the policy provisioning and activation procedures is important in order to minimize client implementation variants.

The section below, after definitions of some key terms, illustrates how policies are provisioned and activated, while the material following gives one example of how these policies can be enforced. In these sections, 5GS terms are used, but the solution is also applicable to EPS, when using eNB/P-GW/PCRF instead of gNB/UPF/PCF.

Policy provisioning and activation

An application client in the vehicle system, which can, for example, be an HD map application, determines the need for fetching the next HD map tile, depending on the current vehicle location, vehicle speed and vehicle route.

The 5G media streaming application function authorizes application policy requests. A mobility service may have the choice of different policies and the application decides, based on need, which policy to activate for the upcoming transaction. While it is specified in the context of media streaming, it is applicable for any traffic type. Different collaboration options are supported. The 5GMS AF is operated by the MNO (trusted AF, Figure 53) or by the MSP (external AF, Figure 54).

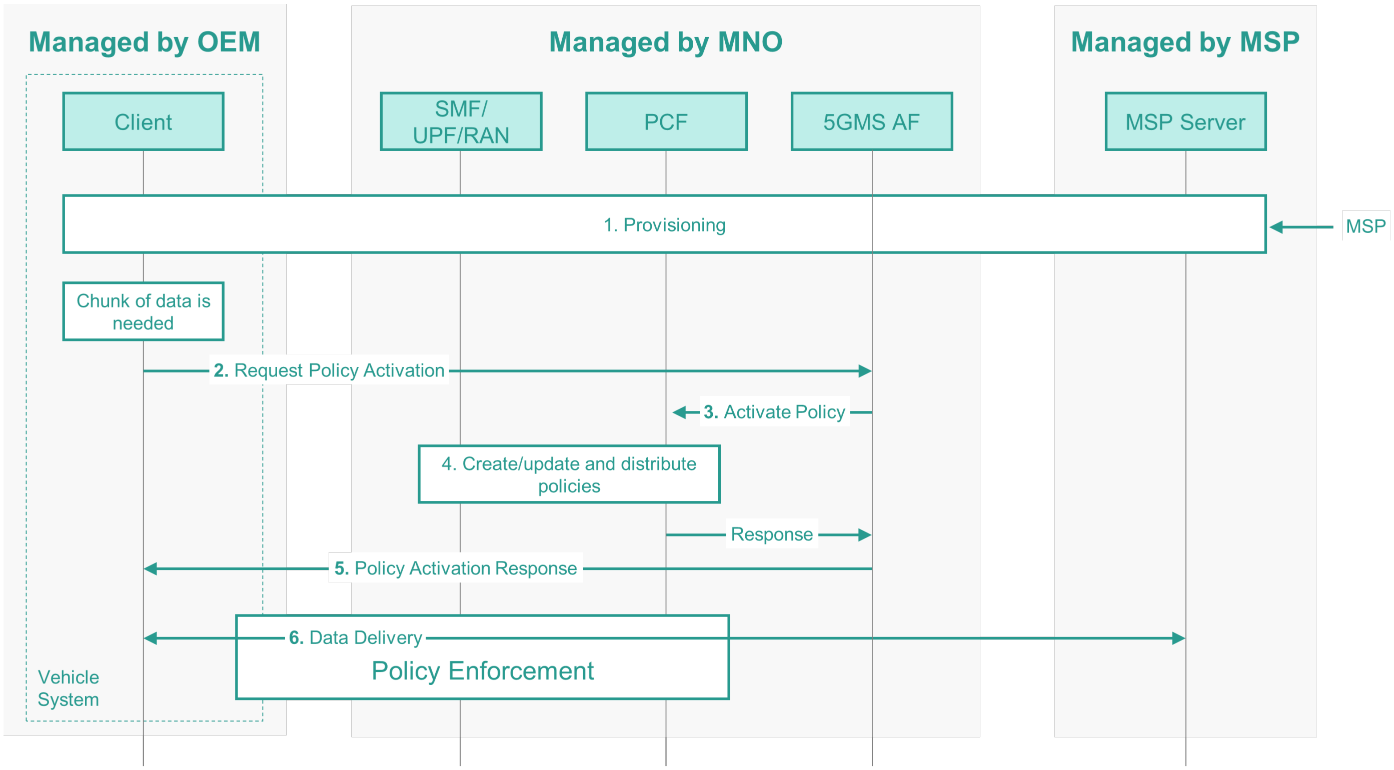

Figure 53. Dynamic policy adaptation – a 5GMS-AG managed by the MNO.

The workflow is as follows:

1) The MSP agrees with a mobile network operator (MNO) to use a set of different policies. The policies to be used can be anchored directly as part of an SLA or, in a more modern approach, using dynamic provisioning mechanisms and exposure APIs. As a result, the mobility service instance has a list of policies to choose from and is authorized to trigger them.
2) The client detects that a chunk of data should be uploaded or downloaded within a certain time frame (e.g., provided in seconds). The client sends a policy activation request for a certain policy to the 5GMS AF.
3) The 5GMS AF activates the selected policy in the 5GS via the NEF. The NEF interacts with the PCF. If the 5GMS AF is a trusted entity for the 5GS (Figure 53), it can instead interact with the PCF directly, and thus has more possible interactions.
4) The PCF creates/updates and distributes corresponding dynamic PCC rules to relevant SMFs and other network functions. Depending on which type of enforcement the mobile network operator uses, other entities are involved. For enforcement on the MAC layer in the RAN, the gNB and UE are involved as well.
5) The 5GMS AF sends a response, indicating that the policy request was granted, and potentially includes policy constraints.
6) The client and mobility service instance exchange data (upload and/or download), while the mobile network takes care of the policy enforcement.

Steps 2-5 might be repeated during delivery, if the need to adapt the policy arises (e.g., when changing from a lower-than-best-effort policy back to a best-effort or even prioritized policy).

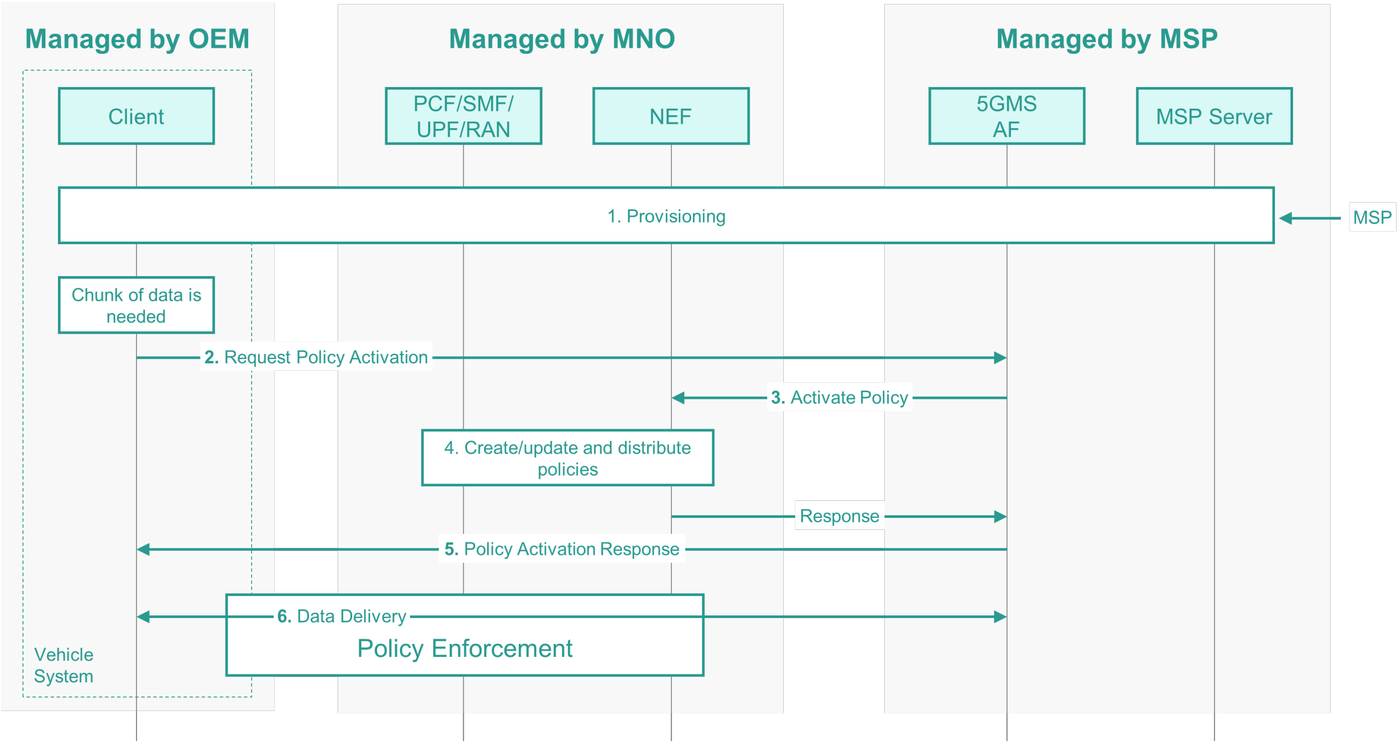

Figure 54. Dynamic policy adaptation – a 5GMS AF managed by the MSP.

Policy enforcement

Several enforcement mechanisms are feasible, and different mobile network operators will have different preferences. Examples would be:

- Using the built-in QoS framework of 4G and 5G networks.
- Using a TCP proxy for traffic shaping.
- Using a QUIC spin bit for pausing the client or server.
- Having a traffic shaper on the IP layer (e.g., leaky bucket).

As one example for policy enforcement, traffic shaping using a transparent TCP proxy is illustrated in Figure 55, where the TCP proxy detects high load conditions (such as by monitoring RTTs) and throttles the corresponding TCP connection. To this end, the TCP proxy would be dynamically configured to adapt read/write operations for reducing throughput. Using this mechanism, the throughput can be throttled to a negligible minimum, by forcing the receive window to zero at the proxy (i.e., stop forwarding any data). In this case, the client sees only a bad connection, and it is up to the client how to react to this (pause delivery, try to reconnect, change access network, etc.). In any case, no specific requirements are put on the client, and it is up to the MNO how to implement the details, without impact on other entities.

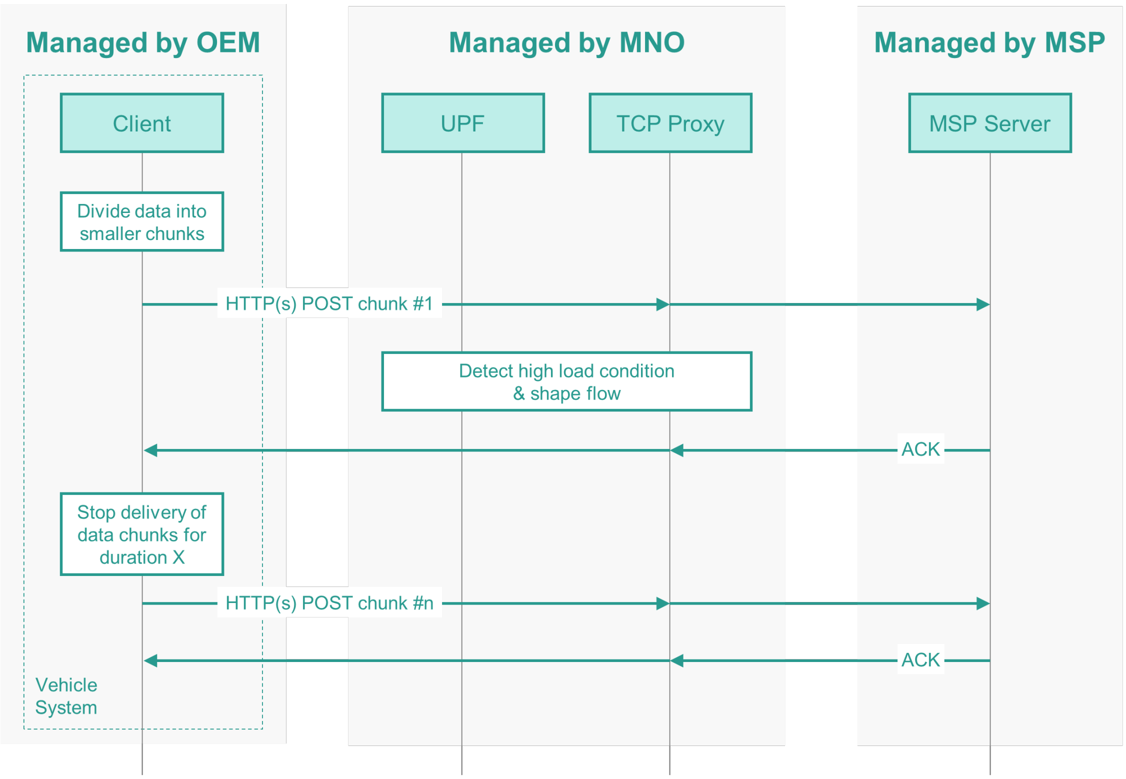

Figure 55. Policy enforcement using a transparent TCP proxy.

3.6.3 Conclusions

The combination of Solutions 2.2 (3GPP UE-based BDT) and 3 (Dynamic Policy Adaptation) is recommended for the future deployment of 5G systems. This solution combination can deal with various requirements on delay tolerance ranging from seconds to days per application. Solution 2.2 and Solution 3 provide the ability for MNOs to dynamically configure policies for UE and easily differentiate between applications. They also provide flexibility for MNOs to optimize efficiency by considering radio resource availability and network conditions. While these solutions provide the most appropriate functionality, their adoption is currently limited; however, since it is targeting the new deployment of 5G systems with less impact on UE and RAN, and because policy negotiations between MSPs and MNOs are based on standardized interfaces, adoption should not be a problem. Furthermore, Solution 3 can be used to support opportunistic data transfer over updated EPS deployments as well, including scenarios with mixed 4G/5G connectivity.

In the case of existing deployed EPS, industry adoption and solution impact on UE and RAN will play a significant role. The combination of Solutions 1.1 (Application-specific Congestion Control for Data Communication (ACDC)) and 2.1 (3GPP network-based Background Data Transfer (BDT)) is recommended to provide opportunistic data transfer on PDN connections of specific APNs pre-defined by MNOs. The benefit of this combination is its high industry adoption, which implies less deployment effort. However, it comes with certain limitations on flexibility for controlling data transfer per application (by mapping application flows to APNs in the vehicle system) and optimizing network resources.

Solution 1.1 can only control data transfer based on a limited number of preconfigured application categories, and Solution 2.1 policy enforcement at the network side is not as resource efficient as Solution 2.2. Consequently, upgrading deployed EPS to adopt Solution 2.2 and Solution 3 is recommended in the long run.

3.7 Service Continuity

3.7.1 Key Issue

In the AECC distributed computing architecture, a vehicle system may be served from different mobility services and IP anchors. Handovers between mobility service instances and/or IP anchors may be triggered by vehicle system movement. The AECC system should be defined such that adequate service continuity can be preserved during such handovers. Service continuity does not mean guaranteeing connectivity between the vehicle system and the mobility service instance, but smooth data transfer between the mobility service instance and the vehicle system can be maintained despite connectivity interference. Although service continuity does not have a direct impact from the disruption of the communication layers' connectivity, service continuity may still be affected in ways listed below.

Change in the IP anchor

A cellular network or WLAN may be used as an access network between the vehicle system and the mobility services. Mobility may cause handover to occur between different access networks, or within the same access network. Vehicle handover to a new IP anchor point will change the vehicle system’s IP, and may cause the mobility service instance to be unable to address the vehicle, resulting in interrupted service. Access network IP anchor handover may need to be handled differently in the cases where the vehicle has a single network interface or multiple network interfaces.

Change of mobility service instance

In this case, it is assumed that a more appropriate mobility service instance is available; most likely this is triggered by requirements known by the application. This scenario applies for a managed mobility service instance handover, where it is assumed that the communication to the new instance is established prior to the mobility service instance handover. The AECC system needs to allow information sharing between AECC system entities to allow selection of the new mobility service instances. Without any handling for mobility service instance handover, service interruption may occur and lead to temporary service blackout.

Figure 56 depicts the process of IP anchor handover (Part 2a) and mobility service instance handover (Part 2b). The order in which those two are executed is not fixed. As a result, a new mobility service instance is used, and it is reached through a new IP anchor (Part 3).

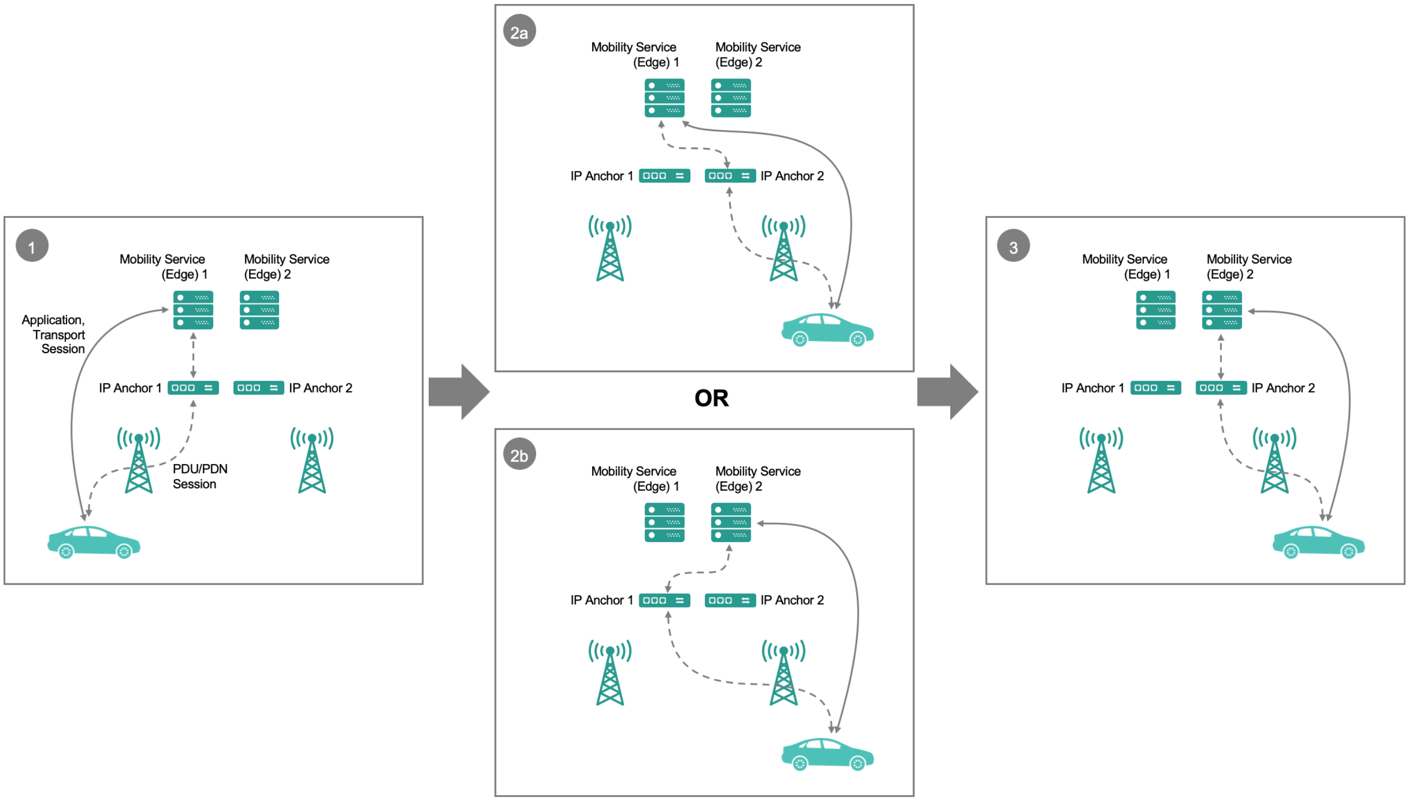
2a
Mobility Service (Edge) 1
Mobility Service (Edge) 2
IP Anchor 1
IP Anchor 2
1
Mobility Service (Edge) 1
Mobility Service (Edge) 2
Application, Transport Session
IP Anchor 1
IP Anchor 2
PDU/PDN Session
OR
3
Mobility Service (Edge) 1
Mobility Service (Edge) 2
IP Anchor 1
IP Anchor 2
2b
Mobility Service (Edge) 1
Mobility Service (Edge) 2
IP Anchor 1
IP Anchor 2

Figure 56. Different orders of mobility service instance and IP anchor handover

This key issue is related to the existing key issues 3.1 Edge Data Offloading and 3.2 Mobility Service Instance Selection. However, this key issue only considers the case where handover of both mobility service instance and IP anchor occurs.

3.7.2 Potential Solutions

The following solutions are described for this key issue.

- Solution 1 – Application Handling of Network Exceptions
- Solution 2.1 – Session and Service Continuity (SSC) mode 3 without application notification
- Solution 2.2 – Session and Service Continuity (SSC) mode 3 with application notification
- Solution 3.1 – Mobility service instance handover before IP anchor handover
- Solution 3.2 – Mobility service instance handover before IP anchor handover with QUIC or Multipath TCP

Many different degrees of freedom exist and the set of solutions below is not exhaustive. Further combinations might exist as well as slightly different variations for each solution. One distinctive element among the solutions is whether break-before-make or make-before-break IP anchor handover is used. These two kinds of handovers are explained under the key issue of edge data offloading. where corresponding SSC modes are introduced. Except for SSC mode 3, equivalent solutions exist for EPS, where SSC mode 1 corresponds to vehicle system-triggered IP anchor handovers and SSC mode 2 corresponds to SIPTO above RAN allowing network-triggered IP anchor handovers in EPS.

Furthermore, some solutions consider that the vehicular router informs the services running in the vehicle system about IP anchor handovers. The other direction, where services inform the vehicular router about mobility service instance handovers, is not considered. The reason is that the vehicle router would not know what to do with that information, as there are typically many different services running in the vehicle system, and a mobility service instance change of one of them would typically not motivate an IP anchor handover affecting all services.

The first set of solutions assumes that the IP anchor handover occurs before any mobility service instance handover. The following set then discusses the case that at least some of the services running on the vehicle system executed a mobility service handover before the IP anchor handover.

Finally, some special features such as using QUIC or Multipath TCP are discussed, along with the benefit they can bring on top of previously discussed solutions.

The first solution description is exhaustive, as it introduces the many different failures that can occur for different kinds of communication patterns. Those remain valid for the solutions following, if not mitigated by the solution, but the detailed descriptions are not repeated.

3.7.2.1 Solution 1 – Application Handling of Network Exceptions

For the first solution it is assumed that the IP anchor handover occurs before the mobility service instance handover. The mobility service instance handover is then assumed to happen based on the solutions described in mobility service instance selection. The IP anchor handover can be network-triggered (SIPTO above RAN for EPS, SSC mode 2 for 5GS) or vehicle system-triggered, initiated by the vehicular router.

Many scenarios, depending on the state the service was in when the IP anchor handover occurred, are possible. From the network (IP) layer perspective, the following might happen:

- The services within the vehicle system did not notice the IP anchor handover and the fact that the external IP address of the vehicle router changed, and with this, likely also the NAT-table in that router was cleared. Any data received at the mobility service, regardless of whether it is in the old or new one, carries the external IP address of the carrier-grade NAT belonging to the new IP anchor.
- Any attempts to send data from the mobility service to the vehicle system will fail, as they will carry the external IP address of the carrier-grade NAT of the old IP anchor. Data will either be dropped there, as the NAT-table entries were removed when the router in the vehicle system was disconnected, or forwarded to the router that is no longer served by the IP anchor and therefore not reached.
- For three different communication patterns in the application layer, this means:
- In the case of a request/reply communication pattern, the IP anchor handover could occur during a transaction. An ongoing upload or download could be interrupted. The service on the vehicle system would eventually get an error message that the transmission cannot be finished. For uploads, the TCP

protocol would eventually report to the application that it cannot complete the transmission. This can take many seconds and depends on how the maximum number of retransmission attempts on timeout is set. For downloads, the error can only be detected within the timeout configuration of the HTTP client, as the TCP protocol cannot know whether the transmission broke or just ended.

In the case of IP anchor handover, when a request was sent but a corresponding reply was not received, the HTTP transaction would eventually time out.

In the case where the application state was handled through cookies and/or unique identifiers, a new mobility service would not contain the state the old one had. The same situation applies for any files that were uploaded.

- In the case of a publish/subscribe communication pattern, for messages sent from the vehicle system the same applies as for uploads using the request/reply communication pattern: the TCP protocol will try to deliver the message until reaching the maximum number of retransmission attempts and then report an error to higher layers. For downlink message delivery the same would occur on the mobility service side. But the problem now is that the service on the vehicle system site remains unaware that the connection is broken and it is not receiving messages intended for it. It would only notice it if keep-alive-messages are enabled for the TCP connection and/or the publish/subscribe protocol (e.g., MQTT) and eventually time out. This typically takes many seconds or even minutes.
- For the combination of both, the following cases are possible: an IP anchor handover during the upload results in the situation described under the first bullet point. Eventually, the service receives an error as the HTTP/TCP protocols are not able to conduct the upload or no reply indicating success is received.

When the IP anchor handover occurs after successfully completing the upload and receiving the reply, the downlink situation of the second bullet point occurs. When the analysis is done, the mobility service fails to reach the service in the vehicle system to inform it about this. The service in the vehicle system keeps waiting as it assumes the analysis is not done. Eventually, a timeout might occur, but it must be set to a value higher than the expected worst-case duration of the analysis.

In the case of an IP anchor handover after receiving the message that the analysis is done but before obtaining the result or while obtaining the result, the description for downloads in the first bullet point applies. In any case the service on the vehicle system site eventually realizes that the download fails.

For publish/subscribe communication patterns, as described in the first bullet above, retrying usually allows normal operation to resume. Between the failed attempt and the successful one, usually a change of mobility service instance is done; for example, because DNS is used for mobility service discovery and the second attempt uses the new mobility service instance address. Application-specific solutions are needed in case state was handled through cookies or unique identifiers.

For publish/subscribe communication patterns, in uplink the broken connection would eventually be detected and the service would be reset, starting by subscribing to topics at the new mobility service instance. In the case where there is no uplink communication, the service has to rely on TCP connection or application timeouts from not-received keep-alive-messages. If this mechanism is disabled, the broken connection might remain undetected. It might in any case result in a long period where messages from the mobility service are not delivered, as keep-alive-timeouts are usually many seconds or minutes.

In the case of mixed communication patterns, different solutions exist:

- The whole analysis is restarted with the new mobility service instance. Where there is IP anchor transition during the upload phase, this is the only option. When done in later phases, it is a waste of resources as the analysis is being done by the old mobility service instance, but the result is not received, as the mobility service instance cannot reach the vehicle system to indicate that the analysis is done.
- The vehicle system could detect that there was a change of mobility service instance, such as that it resubscribed for publish/subscribe services. It would notice that it is still awaiting information when a pending analysis is done, and it might also subscribe to the old mobility service instance to indicate where it can now be reached. This is usually counteracted by the mobility service discovery solutions in place. The vehicle system would only know the mobility server through a Fully Qualified Domain Name (FQDN) that always resolves to the new one once the IP anchor was handed over. Special solutions would also be needed to still reach the old one.
- The old mobility service instance could somehow learn from the new one that the vehicle system is now connected to the new mobility service instance. The old mobility service instance forwards the analysis result to the new mobility service instance, where it is indicated and delivered to the vehicle system in the usual way.

3.7.2.2 Solution 2.1 – Session and Service Continuity (SSC) mode 3 without application notification

SSC mode 3 prevents the total loss of IP connectivity encountered for break-before-make IP anchor handover. As of the writing of this paper, we are not aware of any implementations of SSC mode 3, so we can only make assumptions about how the router in the vehicle system implements SSC mode 3. From experience with using two different APNs/DNNs simultaneously, it can be assumed that the router would add a second external IP interface. All existing NAT table entries would remain associated with the old IP address. Any new NAT table entries would be associated with the new external IP interface. This leads to the following:

- For a request/reply communication pattern, whether download or upload, it depends on the “persistent connection” option. When enabled, the same TCP connection is reused for every request/reply transaction and kept open between transactions. As a result, the new IP interface would never be used. When disabled, the intended behavior is achieved. Ongoing request/reply transactions are completed through the old IP interface and the ones following, assuming that mobility service discovery is triggered before the request, are then done through the new IP interface. Still, extra service-specific handling is needed if the server state was handled through cookies or unique IDs.
- For a publish/subscribe communication pattern the TCP connection is always kept open. The new IP interface would therefore never be used, and traffic would always go through the old IP anchor.
- The combination of both could eventually use the new IP interface for the request/reply portion, if “persistent connection” is disabled in the HTTP client. But the publish/subscribe connection would never transition to the new IP interface in any case.

3.7.2.3 Solution 2.2 – SSC mode 3 with application notification

This is identical to Solution 2.1, but in addition the vehicular router would inform the network that an IP anchor handover took place. Applications using the request/reply pattern should close open TCP connections after completing ongoing request/reply transactions, or immediately, if none are ongoing. Publish/subscribe communication pattern services should be reset such that all subscriptions are redone before the TCP connections are closed.

At present, no standardized methods exist for how the vehicular router could inform other applications in the vehicular system about the IP anchor handover.

3.7.2.4 Solution 3.3 – Mobility service instance handover before IP anchor handover

Although many different services would be running in the vehicle system, this solution assumes that, at least for some of them, a mobility service instance handover occurs before the IP anchor handover. This would require using mobility service instance selection procedures that do not select the mobility service instance based on the IP anchor that is being used. We would still assume that eventually an IP anchor handover is triggered. Without this assumption, the scenario would be very similar to the one solved by the mobility service instance selection solutions.

As the services trigger the mobility service instance handover, it can be assumed that this is done seamlessly: connections are (re)established as needed and subscriptions for publish/subscribe services renewed. Furthermore, services should be designed in a way to prevent mobility service instance handovers breaking an ongoing communication transaction. For the request/reply communication pattern, none of the mobility service instance selection solutions can cause this. Also, for the publish/subscribe communication pattern, it can be assumed that even for the “Mobility Service Assignment by a Selection Function” solution for mobility service instance selection, corresponding services in the vehicle system would either first subscribe to the new mobility service before releasing the old one, or at least indicate to the old one that they cannot be reached anymore, by unsubscribing.

For the combination of request/reply and publish/subscribe communication patterns, there is a risk that the mobility service instance handover occurs before the processing is done and indicated to the vehicle system. This should be solved by either preventing mobility service instance handovers while waiting for processing results or assuring that the results can be delivered through the new mobility service instance.

Once the IP anchor handover is executed, the same connection interruptions described under Solution 1 will occur if SSC mode 3 (Solution 2.1 and Solution 2.2) is not used. In this case, the recovery mechanisms described under Solution 1 should handle the network exception, with the difference that within the recovery procedure no mobility service instance handover is needed. It therefore corresponds to normal recovery mechanisms applied for any connection interruption and can be considered common practice in distributed software development.

SSC mode 3 without application notification (Solution 2.1) will result in the same situation as described above, as the services running in the vehicle system would not notice the IP anchor handover and would mostly not behave in a way that enables the vehicular router to eventually route their traffic through the new IP interface and therefore new IP anchor. Exceptions exist, as described under Solution 2.1. In the case of SSC mode 3 with application notification (Solution 2.2), services running on the vehicle system can take actions to ensure that their traffic is eventually routed through the new IP interface and IP anchor. Challenges described under Solution 2.2 remain valid.

3.7.2.5 Solution 3.2 – Mobility service instance handover before IP anchor handover with QUIC or Multipath TCP

All previously described solutions require TCP connections to be re-established. The reason is that upon IP anchor handover IP address changes also occur, while IP addresses are used in legacy TCP to identify connections. QUIC and Multipath TCP use unique identifiers instead of IP addresses.

It is assumed that none of the solutions described below work with SSC mode 3 as they do not re-establish TCP connections but seamlessly resume them. The vehicular router likely relies on TCP connections to be re-established in order to route them through the new IP interface and with that the new IP anchor. This is speculative, as SSC mode 3 implementations and corresponding routers currently do not exist.

After IP anchor handover, a request/reply communication type of service would resume normal operation. The server would detect that traffic is now coming from a different IP address and would update the corresponding IP address for the connection that it identifies through the unique identifier.

For publish/subscribe services, there must be a notification from the vehicular router to the corresponding service. Upon receiving this notification, the client must send any kind of data through the TCP connection to trigger an IP address update for that TCP connection at the mobility service. Alternatively, keep-alive messages sent by the client can be used, but this could delay the update process. As long as the update is not done, the vehicle system will not receive downlink messages intended for it, as they are sent to the wrong, old IP anchor.

Services applying a combination of request/reply and publish/subscribe communications will benefit from the seamless TCP connection update for the request/reply part and will have to ensure that the publish/subscribe part is proactively updated, such as by simply re-subscribing, to make sure that the information about finished processing reaches the vehicle system. This only applies if the IP anchor handover occurs after uploading the data to be processed but before the notification that the processing was completed was received. The vehicular router must inform such services about an IP anchor handover so that they can act accordingly in case there are pending notifications.

3.7.3 Conclusions

The recommendations for the mobility service instance selection solutions should be followed as prerequisites, as they cover the mobility service instance handover, while the following focus on the IP session anchor handover. Solution 1 is the preference for now, as any software intended for wireless networks should include procedures to recover from broken connections.

Once SSC mode 3 is deployed AND solutions are in place for the router to inform application clients, Solution 2.2 should be considered. Solution 2.1 will reduce the service interruption time, but at some point there will be an interruption nonetheless. Solution 2.1 might reduce this interruption, but the reduction is likely negligible compared to the ones that will still happen.

Solution 3.2 can be considered for selected applications, and it offers a truly uninterrupted experience when combined with Solution 2.2. Solution 3.1, as with 2.1, adds complexity but only partially prevents breaking of connections. It should therefore not be pursued.

3.8 Geolocation Services

3.8.1 Key Issue

Geolocation services are used in mobility applications to pre-process data from vehicle and other systems to detect changes such as lane closures or new road-signs, hazards such as debris or black ice, and dynamic conditions such as surges or roadwork, in a given geospatial area. Geolocation processing is split among agents by areas for scale across edge locations, and the associations between vehicles and these agents are not random but based on the areas vehicles are in. The dynamics in which agents move between edge locations based on vehicle density and the dynamics in which vehicles drive across areas cause the key issues.

- Dynamic resource allocation: In a metropolitan area, there may be millions of vehicles providing raw geospatial data and subscribing to geolocation topics. However, the geographic distribution of vehicles in use changes over time, so it is necessary to allocate resources to the geolocation agents elastically on the fragmented infrastructure of the edge.
- Protection of vehicle system privacy: The association between vehicles and geolocation services partitioned by agents enables unwanted tracking of the vehicles by listeners in the middle even if communication is encrypted.
- Seamless switching between geospatial contexts: The geolocation agent processes the geospatial data sent from vehicles, as an essential element of a mobility service that provides geolocation information. This means that a given vehicle may be the source of data for one agent one minute and another agent the next.
- Continued scale of notifications: Given a set of thousands of geolocation notifications and millions of vehicles that change their location each second, the network needs to calculate which vehicle needs to receive which notification using what network provider-assigned IP address.

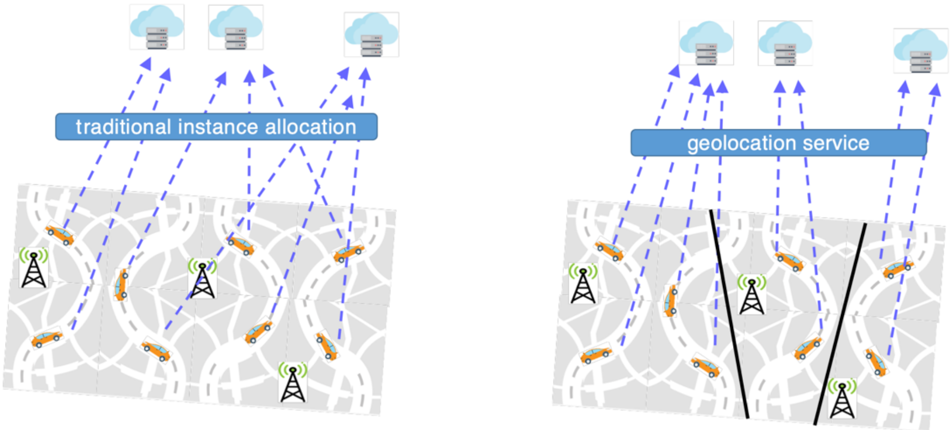

Figure 57. The left side of the image depicts a traditional centralized cloud with shared data fabric instance allocation, in comparison with the requirements of geolocation services distributed at the edge, shown at right.

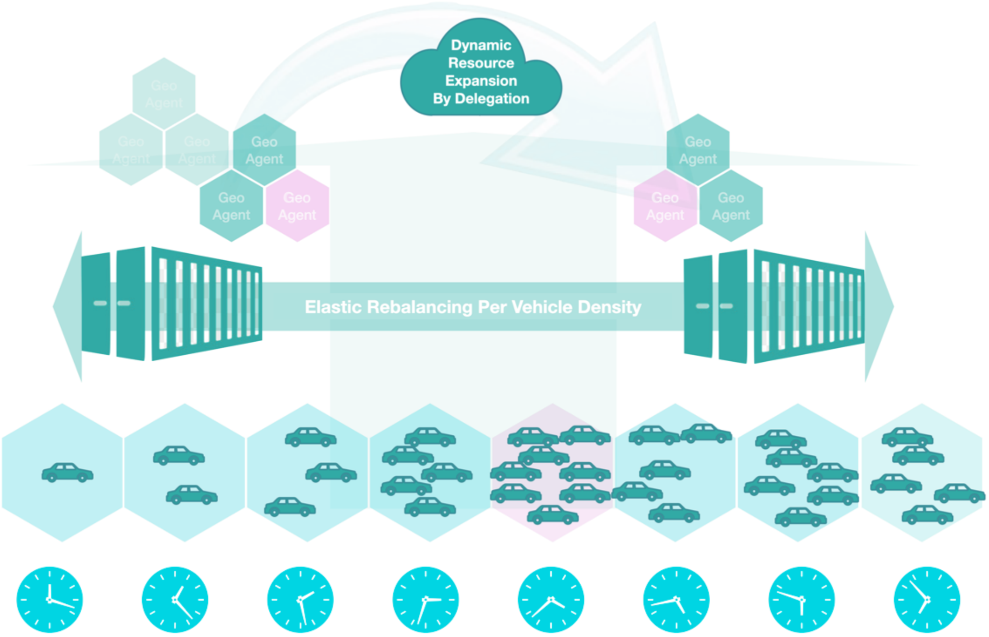

Figure 58: Moving vehicles' dynamic geolocation association (context, privacy, notifications) and portable agents for dynamic allocation.

As can be seen in Figure 58, the number of geolocation areas (agents) is the same throughout the day, as at any given moment uploads relating to any area may need to be contextually processed by the area agent. However, the density of vehicles changes during the day, which means that the same geolocation agents occupy more or less of the edge resources based on needs. Therefore, to optimize costs and service availability, agents are delegated between servers and edge locations such that there are fewer agents per server during peak hours. This portability can confuse the vehicles interacting with geolocation agents.

As can be seen, vehicles need to constantly context-switch their agent interactions while they drive across locations, which is not trivial to do continuously without interruption. A given vehicle functions as the sensor of one geolocation agent one second and a different one the next.

The interaction between vehicles and agents needs to be protected from being tracked by any listener in the middle. Even encrypted messages disclose the location of each vehicle during the day. The mobile edge geolocation network needs to scale for dynamic notifications.

Finally, given a set of thousands of geolocation situations at any moment and millions of vehicles that change their location each second and may also change their IPs while moving between access anchors, the network needs to calculate which vehicle needs to receive which notification using what network provider-assigned IP address right now.

3.8.2 Potential Solutions

As described above, geolocation processing is split among agents by areas for scale across edge locations, and the associations between vehicles and these agents are not random but based on the areas vehicles are in. The dynamics in which agents move between edge locations based on vehicle density and the dynamics in which vehicles drive across areas cause the key issues. To solve these, we define a dataflow virtualization layer (DFV) between moving vehicles and moving agents. However, as there are several possible industry patterns for DFV solutions as described below, it is necessary to analyze and compare them to determine the best architecture.

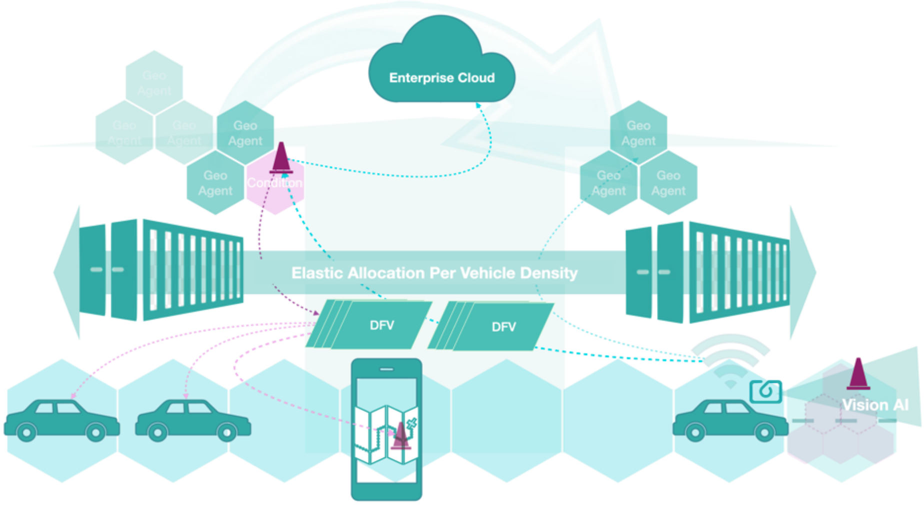

Figure 59: Dataflow virtualization (DFV) between moving vehicles and portable agents.

The following DFV solutions are described:

- Solution 1: Off-path DFV
  - Solution 1.1– DNS CDN for resolving well known geolocation domain names to Ips
  - Solution 1.2 – HTTP redirect resolving well-known geolocation URLs to actual IPs
- Solutions 2: On-path DFV
  - Solution 2.1 – Underlay anycast translating well-known geolocation IPs to actual IPs
  - Solution 2.2 – Overlay EID “map & encap” geolocation IPs in underlay actual IPs

Portable and distributed geospatial agents can be used to address the dynamic nature of geolocation services in an open environment like a city. While the number of geospatial areas in a city rarely changes, the amount of processing per area varies considerably during the day. Portable and distributed implementations of geospatial agents enable elastic delegation across computing platforms based on activity levels. When activity levels are lower, more agents can be hosted on fewer computing platforms, and when activity levels increase, the same geospatial agents can be rebalanced across more computing platforms.

Portability of a geospatial agent mainly involves the portability of its addressable queues and channels from and to vehicle systems. Geolocation functions are pervasive, and the working state quickly regenerates on any edge computing location. The portability of geospatial agents’ queues and channels mainly involves some form of dataflow virtualization between the addressable entities on computing platforms and on vehicle systems. Vehicle systems can also be regarded as addressable entities as far as DFV is concerned, in order to address geoprivacy and geospatial topic subscription continuity key issues.

DFV enables the rebalancing of geolocation processing agents across varying edge computing locations based on road activity without major interruption to the service. Rebalancing also occurs if edge computing locations suffer outages or are disconnected from the network. DFV also simplifies geolocation context switching by vehicle systems driving in and out of each area. DFV, when applied to the virtual addressing of vehicle systems, in addition to geolocation processing agents, solves geoprivacy and subscription continuity. Depending on the DFV solution chosen, it can also solve subscription notification scaling. The industry distinguishes between two categories of DFV solutions: off-path and on-path.

3.8.2.1 Solution 1.1 – Off-path DNS CDN

The network for mobility services maintains a high-capacity DNS CDN cluster at the edge. The CDN is constantly updated with logical domain names to IP mappings. Clients frequently resolve these geolocation domain names as they drive through geospatial domains, ensuring that the geolocation agent for the geospatial domain they are in has not been rebalanced.

3.8.2.2 Solution 1.2 – Off-path HTTP redirect

The network for mobility services maintains an HTTP redirect cluster. Randomly picked redirectors steer geolocation URL requests back to clients with the current IP locations of the geolocation agents. Clients then resend HTTP requests to the correct IP location.

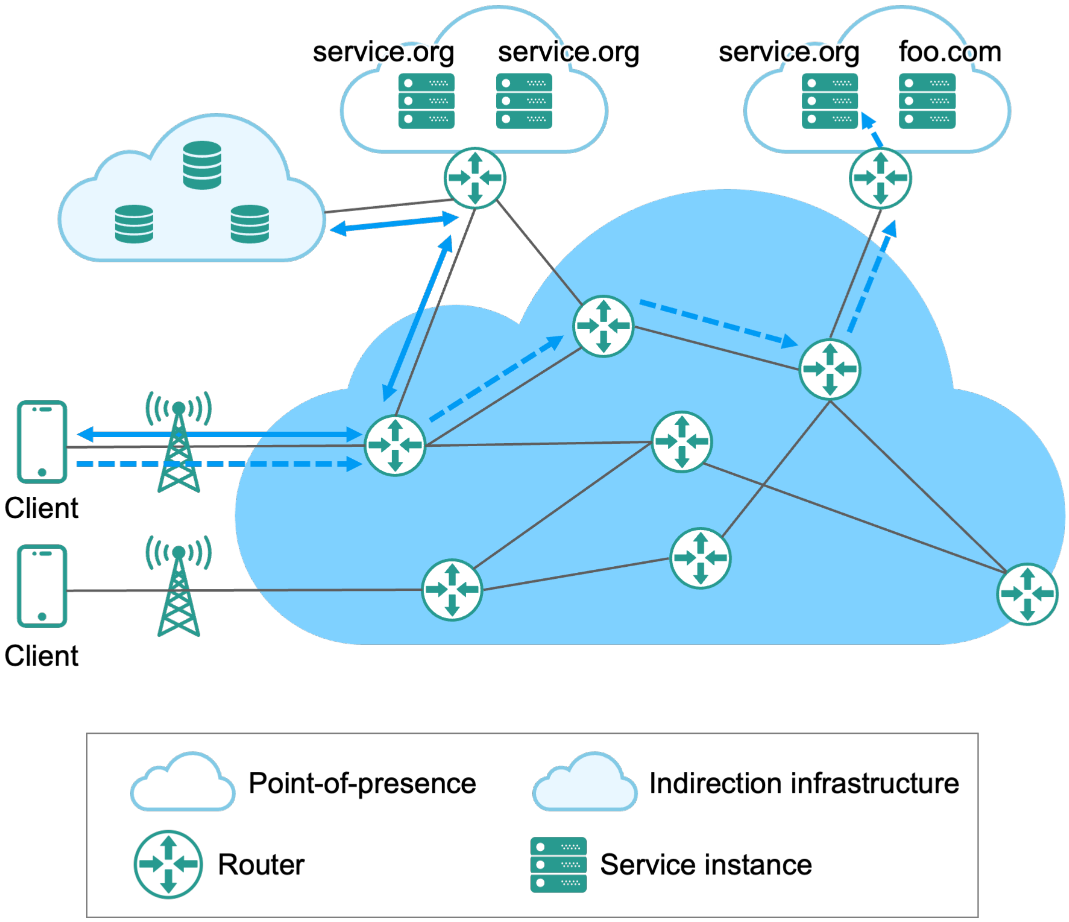

Figure 60: Off-path solution for DFV.

3.8.2.3 Solution 2.1 – On-path underlay (BGP-anycast/dyncast)

Clients use well-known IP addresses for their requests; designated BGP aggregators perform network address translation (NAT) and switch the well-known IP headers with current IPs of the geolocation agents based on NAT tables updated at these BGP gateways.

3.8.2.4 Solution 2.2 – On-path overlay (LISP/SDWAN-EID)

The network for mobility services distributes tunneling gateways across the edge connected by regular IPs. Geolocation agents are associated with these gateways based on current rebalancing. Clients then communicate with agents via any one of the tunnel gateways using logical geolocation IPs and logical ephemeral IPs endpoint identifiers (EID). The gateways share a mapping service.

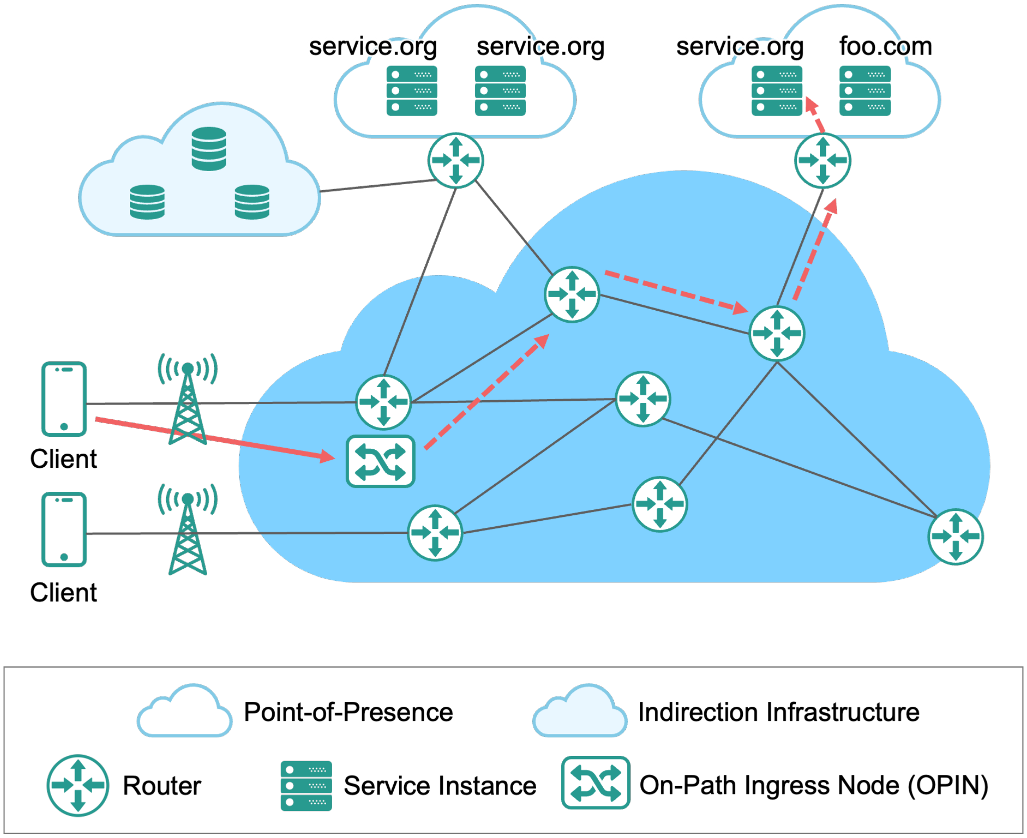

Figure 61: On-path solution for DFV.

3.8.2.5 Solutions Summary

As discussed, the root cause of geolocation key issues is in the associations between vehicle systems and geolocation edge processing agents. The changing number of active vehicles in geolocation areas, the interaction and context switching between vehicle systems and these agents, and the mere dynamics of vehicle systems' geolocations and connectivity cause these key issues. Virtual addressing of vehicle systems and of geolocation agents and a virtualized dataflow (DFV) between them solves these key issues. DFV addresses the key-issues stemming from the dynamic association between moving vehicle systems and moving geolocation agents. Geospatial addressing of agents allows them to move and rebalance across computing locations, and allows for simple context switching for vehicles driving through geospatial areas, ephemeral assigned addressing of vehicle systems maintain subscriptions across IP anchors, they are used to track and scale notifications while preserving geoprivacy.

The advantage of on-path solutions is that they apply dynamic mappings at the network layer and provide built-in security and privacy measures. On-path solutions work well in decentralized edge environments because they allow for dynamic instance selection based on real-time traffic and resource availability. The challenge of underlay-based solutions such as BGP anycast in geodistributed edge environments is the complexity and overhead of maintaining routing tables and the potential for routing loops and blackholes. LISP and similar session-aware overlay networks such as WireGuard do not suffer from these risks, but they add an IP in IP VXLAN or similar encapsulation header overhead to all virtualized communications.

The challenge of off-path mechanisms is that they require centralized cloud environments to work well at scale. The cloud hides many of the changes that would otherwise invalidate cached resolutions/redirects by a large number of clients; having no centralized masking can result in mass interruptions and re-resolution storms. Redirects work well within co-located ultra-low-latency cloud datacenters. However, they tend to oscillate and produce non-deterministic latencies across edge locations. This is a result of percolation by mass uncoordinated clients (unlike on-path network-aware aggregation) leading also to slow recovery from disconnects.

3.8.3 Conclusions

In conclusion, dataflow virtualization addresses the key issues stemming from the dynamic association between moving vehicle systems and moving geolocation agents by allowing for geospatial addressing of agents and ephemeral addressing of vehicles. Logical addressing enables dynamic resource allocation, simple context switching for vehicles, preservation of geoprivacy, and scaling subscription tracking and notifications. There are two main categories of DFV solutions: off-path and on-path techniques. Off-path solutions are simpler but resolve over mobile access to millions of clients, and are therefore more suitable for centralized clouds, which hide the need for most resolutions with load-balancers. On-path solutions require data-path processing but are better for distributed environments, and they resolve inline without involving mobile access and mass clients. A combination of off-path and on-path solutions may also be used when edge distribution is limited and resembles a single global metro area cloud. Out of the on-path solutions, overlay-based traffic steering, such as LISP-EIDs or WireGuard, is the most flexible and least intrusive compared to BGP anycast, and it scales DFV for both upstream uploads and downstream notifications while protecting privacy.

4 A Path Forward for New Solutions

The sets of solutions for the key issues presented in this version of this AECC technical report further detail the path on how to drive data to the edge from an end-to-end networking perspective. The AECC believes that this proposal should play a significant role in supporting new automotive service scenarios.

In the future, the AECC will continue to identify new key issues and investigations into solutions around the end-to-end networking aspect to meet growing data volumes with high demands on the security, sovereignty and efficiency of the delivered data of the automotive industry.

References

[1] GSMA, 2025 Every Car Connected: Forecasting the Growth and Opportunity. Feb. 2012. https://www.gsma.com/iot/wp-content/uploads/2012/03/gsma2025everycarconnected.pdf

[2] Ericsson, Ericsson Mobility Report, June 2019 edition. https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf

[3] Ericsson, Ericsson Mobility Visualizer, June 2019. https://www.ericsson.com/en/mobility-report/mobility-visualizer?f=14&ft=2&r=1&t=21,22,23&s=14&u=1&y=2018,2024&c=1s

[4] Cisco, Cisco Annual Internet Report (2018–2023) White Paper. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf

[5] Automotive Edge Computing Consortium, White Paper, General Principle and Vision, Version 3.0.0, Jan. 2020. https://aecc.org/wp-content/uploads/2020/07/General_Principle_and_Vision_January_31_2020.pdf

[6] 3GPP, Technical Specification 23.401: General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, Version 16.3.0, Jun. 2019. http://www.3gpp.org/ftp//Specs/archive/23_series/23.401/23401-g30.zip

[7] 3GPP, Technical Specification 23.501: System Architecture for the 5G System, Version 16.1.0, Jun. 2019. http://www.3gpp.org/ftp//Specs/archive/23_series/23.501/23501-g10.zip

[8] oneM2M, Technical Specification 0001: Functional Architecture, Version 4.1.0, Jun. 2019. http://member.onem2m.org/Application/documentapp/downloadLatestRevision/default.aspx?docID=30069

[9] 3GPP, Technical Specification 23.402: Architecture enhancements for non-3GPP accesses, Version 15.3.0, Mar. 2018. http://www.3gpp.org/ftp//Specs/archive/23_series/23.402/23402-f30.zip

[10] 3GPP, Technical Specification 24.312: Access Network Discovery and Selection Function (ANDSF) Management Object (MO). Version 15.0.0, Jun. 2018. http://www.3gpp.org/ftp//Specs/archive/24_series/24.312/24312-f00.zip

[11] 3GPP, Technical Specification 24.302: Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP Access Networks; Stage 3. Version 16.0.0, Mar. 2019. http://www.3gpp.org/ftp//Specs/archive/24_series/24.302/24302-g00.zip

[12] IETF, Internet-Draft: Generic Multi-Access (GMA) Convergence Encapsulation Protocols. https://tools.ietf.org/html/draft-zhu-intarea-gma-03

[13] IETF, RFC: Multi-Access Management Services. https://www.rfc-editor.org/rfc/rfc8743.html

[14] 3GPP, Technical Specification 36.300: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2. Version 15.7.0, Sep. 2019.
http://www.3gpp.org/ftp//Specs/archive/36_series/36.300/36300-f70.zip

[15] 3GPP, Technical Specification 33.401: 3GPP System Architecture Evolution (SAE); Security architecture. Version 16.1.0, Dec. 2019.
http://www.3gpp.org/ftp//Specs/archive/33_series/33.401/33401-g10.zip

[16] 3GPP, Technical Specification 36.361: Evolved Universal Terrestrial Radio Access (E-UTRA); LTE-WLAN Radio Level Integration Using Ipsec Tunnel (LWIP) encapsulation; Protocol specification. Version 15.0.0, Jul. 2018.
http://www.3gpp.org/ftp//Specs/archive/36_series/36.361/36361-f00.zip

[17] 3GPP, Technical Report 23.793: Study on access traffic steering, switch and splitting support in the 5G System (5GS) architecture. Version 16.0.0, Dec. 2018.
http://www.3gpp.org/ftp//Specs/archive/23_series/23.793/23793-g00.zip

[18] 3GPP, Technical Specification 37.340: NR; Multi-connectivity; Overall description; Stage-2. Version 15.7.0, Sep. 2019.
http://www.3gpp.org/ftp//Specs/archive/37_series/37.340/37340-f70.zip

[19] IETF, RFC 2132: DHCP Options and BOOTP Vendor Extensions.
https://tools.ietf.org/html/rfc2132

[20] IETF, RFC 5970: DHCPv6 Options for Network Boot.
https://tools.ietf.org/html/rfc5970

[21] 3GPP, Technical Specification 22.011: Service Accessibility. Version 17.0.0, Dec. 2019.
http://www.3gpp.org/ftp//Specs/archive/22_series/22.011/22011-h00.zip

[22] 3GPP, Technical Specification 23.502: Procedures for the 5G System; Stage 2. Version 16.3.0, Dec. 2019.
http://www.3gpp.org/ftp//Specs/archive/23_series/23.502/23502-g30.zip

Acknowledgements

This document is provided by the Automotive Edge Computing Consortium (AECC) Technical Solution Working Group (WG2). The following people have provided valuable contributions to this document:

ORGANIZERS

Leifeng Ruan (WG Chair)

Mikael Klein (WG Vice-Chair)

Lei Zhong (Chief Editor)

CONTRIBUTORS (in alphabetical order)

Lidwina Andarini

Hirochika Asai

Sharon Barkai

Bastian Cellarius

Zongrui Ding

George Ericson

Toru Furusawa

Brian Keller

Seiichi Koizui

Tomoyuki Kurobe

Clara Li

Morgan Lindqvist

Thorsten Lohmar

Takuya Miyasaka

Koya Mori

Joel Obstfeld

Ryokichi Onishi

Ryo Tamura

Xiaopeng Tong

Wenlei Wu

Version History

Date	Version	Description
2019/09	1. 0.0	Initial AECC Reference Architecture, three key issues and corresponding solution recommendations for Edge Data Offloading, Mobility Service Instance Selection and Vehicle System Reachability.
2020/07	2.0.0	Added three key issues: Access Network Selection, Provisioning and Configuration Update and Opportunistic Data Transfer. Added initial Distributed Computing Reference Model into AECC Reference Architecture.
2023/04	3.0.0	Added two key issues: Service Continuity and Geolocation Services. Removed the content on Distributed Computing to a separate technical report.

About the Automotive Edge Computing Consortium (AECC)

The AECC is a consortium of leaders across industries focused on driving the evolution of edge network architectures and computing infrastructures to support high-volume data services in a smarter, more efficient connected-vehicle future. AECC members are key players in the automotive, high-speed mobile network, edge computing, wireless technology, distributed computing and artificial intelligence markets.

Our mission is to develop an open, technology-agnostic framework to support the transfer of data and communications between the vehicle and local networks near the source of the data, and then to a centralized cloud in a seamless, safe, reliable and optimized manner. Our members collaborate on the development of use cases, technical solutions and reference architectures.

- **Use Cases** – Create use cases and requirements in networking and computing for connected services in cars.
- **Technical Solutions** – Provide inputs to standards and open source communities on best practices for deploying distributed and layered networking and computing solutions for connected cars.
- **Reference Architectures** – Develop architectures for next-generation mobile networks and distributed computing that are suitable for automotive use cases.

We invite you to join the AECC and add your insights and influence. Visit https://aecc.org/ to learn more.